PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : C12Q 1/68, C12N 15/10

A1

(11) International Publication Number: **WO 98/14619**

(43) International Publication Date: 9 April 1998 (09.04.98)

(21) International Application Number: PCT/US97/18344

(22) International Filing Date: 3 October 1997 (03.10.97)

(30) Priority Data:
08/723,646 3 October 1996 (03.10.96) US

(71) Applicant *(for all designated States except US)*: INCYTE PHARMACEUTICALS, INC. [US/US]; 3174 Porter Drive, Palo Alto, CA 94304 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: WANG, Bruce, B. [US/US]; 1123 Banyon Way, Pacifica, CA 94404 (US). CHUNG, Alicia [US/US]; 2939 20th Avenue, San Francisco, CA 94132 (US). GUEGLER, Karl, J. [CH/US]; 1048 Oakland Avenue, Menlo Park, CA 94025 (US). YANG, Zhi [CN/US]; 600 Coleman Avenue, Menlo Park, CA 94025 (US). COCKS, Benjamin, Graeme [AU/US]; 4292 D Wilke Way, Palo Alto, CA 94306 (US). STUART, Susan, G. [US/US]; 1256 Birch Street, Montara, CA 94037 (US).

(74) Agent: BILLINGS, Lucy, J.; Incyte Pharmaceuticals, Inc., 3174 Porter Drive, Palo Alto, CA 94304 (US).

(81) Designated States: AT, AU, BR, CA, CH, CN, DE, DK, ES, FI, GB, IL, JP, KR, MX, NO, NZ, RU, SE, SG, US, ARIPO patent (GH, KE, LS, MW, SD, SZ, UG, ZW), European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, ML, MR, NE, SN, TD, TG).

Published

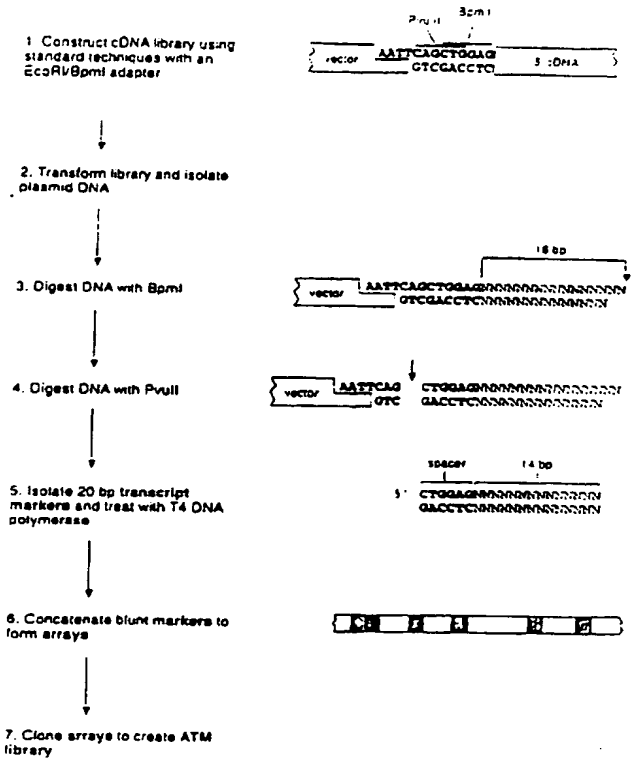
With international search report.

Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.

(54) Title: METHODS FOR GENERATING AND ANALYZING TRANSCRIPT MARKERS

(57) Abstract

The present invention relates generally to the field of molecular biology and specifically to rapid, high–throughput gene discovery methods that facilitate genome closure and to methods for analyzing gene expression patterns. The present invention provides methods and vectors useful for constructing libraries of transcript markers. The present invention also provides sequence specific methods for extending the nucleotide sequence of partial transcripts in a high–throughput manner using polymerase chain reaction.



FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	**ES**	Spain	**LS**	Lesotho	**SI**	Slovenia
AM	Armenia	**FI**	Finland	**LT**	Lithuania	**SK**	Slovakia
AT	Austria	**FR**	France	**LU**	Luxembourg	**SN**	Senegal
AU	Australia	**GA**	Gabon	**LV**	Latvia	**SZ**	Swaziland
AZ	Azerbaijan	**GB**	United Kingdom	**MC**	Monaco	**TD**	Chad
BA	Bosnia and Herzegovina	**GE**	Georgia	**MD**	Republic of Moldova	**TG**	Togo
BB	Barbados	**GH**	Ghana	**MG**	Madagascar	**TJ**	Tajikistan
BE	Belgium	**GN**	Guinea	**MK**	The former Yugoslav Republic of Macedonia	**TM**	Turkmenistan
BF	Burkina Faso	**GR**	Greece			**TR**	Turkey
BG	Bulgaria	**HU**	Hungary	**ML**	Mali	**TT**	Trinidad and Tobago
BJ	Benin	**IE**	Ireland	**MN**	Mongolia	**UA**	Ukraine
BR	Brazil	**IL**	Israel	**MR**	Mauritania	**UG**	Uganda
BY	Belarus	**IS**	Iceland	**MW**	Malawi	**US**	United States of America
CA	Canada	**IT**	Italy	**MX**	Mexico	**UZ**	Uzbekistan
CF	Central African Republic	**JP**	Japan	**NE**	Niger	**VN**	Viet Nam
CG	Congo	**KE**	Kenya	**NL**	Netherlands	**YU**	Yugoslavia
CH	Switzerland	**KG**	Kyrgyzstan	**NO**	Norway	**ZW**	Zimbabwe
CI	Côte d'Ivoire	**KP**	Democratic People's Republic of Korea	**NZ**	New Zealand		
CM	Cameroon			**PL**	Poland		
CN	China	**KR**	Republic of Korea	**PT**	Portugal		
CU	Cuba	**KZ**	Kazakstan	**RO**	Romania		
CZ	Czech Republic	**LC**	Saint Lucia	**RU**	Russian Federation		
DE	Germany	**LI**	Liechtenstein	**SD**	Sudan		
DK	Denmark	**LK**	Sri Lanka	**SE**	Sweden		
EE	Estonia	**LR**	Liberia	**SG**	Singapore		

METHODS FOR GENERATING AND ANALYZING TRANSCRIPT MARKERS

TECHNICAL FIELD

The present invention relates generally to the field of molecular biology and specifically to rapid, high-throughput gene discovery methods that facilitate genome closure and to methods for analyzing gene expression patterns.

BACKGROUND ART

Many genes have been isolated and their structures determined since the introduction of recombinant cDNA technology. It has been predicted that through the efforts of the Human Genome Project, sequencing of the entire human genome (genome closure) will be accomplished sometime between 2001 and 2005 (Boguski, 1995, Molecular Medicine 333:645-647). Strategies for achieving genome closure include methods for constructing 3' directional normalized libraries which equalize cDNA representation (Soares et al. 1994, Proc. Natl. Acad. Sci. 91:9928 and Patanjali et al, 1991, Proc. Natl. Acad. Sci. USA 88:1943-1947) and methods for creating cDNA libraries based on subtractive hybridization techniques which provide methods for isolating the coding sequences of genes which are differentially expressed, such as during development or in disease states which are described in Rubenstein et al (1990, Nucleic Acids Research, 18:4833-4842); Travis et al. (1988, Proc. Natl. Acad. Sci. USA, 85:1696-1700).

Elucidation of gene expression represents another level of complexity equally important to the elucidation of genetic structure. The generation of a gene expression pattern can be used directly as a diagnostic profile or as a gene discovery method. Seilhamer et al. (WO 95/20681, filed January 27, 1995) disclose methods for the high-throughput sequence-specific analysis of cDNAs and generation of transcript images. Matsubara et al (WO 95/14772, filed November 11,1994) disclose methods for generating 3' directed cDNA libraries which accurately reflect the abundance ratio of mRNA in a cell. Velculescu et al. (1995, Science 270:484-487) describe a method for the analysis of gene "tags" or transcripts which uses type IIs restriction enzymes and involves the generation and analysis of short, 3' nucleotide sequences which may inherently contain a substantial amount of 3' non-coding information. Kato (1995, Nucleic Acids Research 23:3685-3690) describes a method for the identification of 3' end cDNA fragments which involves the use of type IIs restriction enzymes and PCR methodology. Kato notes that there are several technical limitations to the method including the presence of PCR generated artifacts and the fact that cDNA sequences lacking enzyme recognition sites will not be displayed.

Aoto et al. (1995, Eur. J. Biochem. 234:8-15) describe isolation of a cDNA clone

obtained from a cDNA library of mRNA prepared from cells treated with 5-azacytidine. The deoxycytidine analog, 5-aza-2'-deoxycytidine (5-azadCyd), is highly toxic in cultured cells and animals and has been used clinically as an anti-cancer agent.

5-azadCyd has been used experimentally as a DNA methylation inhibitor to induce gene expression and cellular differentiation (Juttermann et al. 1994, Proc. Natl. Acad. Sci. USA 91: 11797-11801).

In spite of the availability of methods designed to facilitate gene discovery and elucidate gene expression, there remains a need in the art for methods which will expedite the process of gene discovery in a rapid, high-throughput manner, thereby contributing to the process of gene closure. There also remains a need in the art for methods for analyzing gene expression patterns in a rapid, high throughput manner.

DISCLOSURE OF THE INVENTION

The present invention relates, in part, to rapid, high-throughput gene discovery methods that facilitate genome closure and to methods for analyzing gene expression patterns. The present invention also relates to methods for the rapid, sequence-specific identification of transcripts derived from an mRNA population. The present invention further relates to methods for extending the nucleotide sequences of partial transcripts in a high-throughput manner using polymerase chain reaction technology.

The present invention provides rapid, high-throughput methods for generating and analyzing transcript markers from the 5' most end of cDNAs and methods for generating and analyzing two, discontinuous transcript markers from a single cDNA providing the advantage of obtaining more information from a single transcript than is possible by current methods. In one aspect of the present invention, the two discontinuous markers are derived from both the 5' and 3' ends of a single cDNA and in another embodiment of the present invention, the two discontinuous markers are derived from random areas of the cDNA.

In one aspect of the present invention, a method is provided for generating and analyzing transcript markers from the 5' most end of individual cDNAs of a cDNA library comprising the steps of: obtaining a cDNA library comprising individual cDNAs having a first restriction endonuclease site for a restriction endonuclease that digests the cDNA at the 5' most end within an expected distance from its recognition site and a second endonuclease restriction site; subjecting the cDNAs to digestion with the first and the second restriction endonuclease thereby excising transcript markers from the 5' most end of the individual cDNAs; ligating said transcript

markers to a vector; transforming said vector containing the transcript markers in a host cell; culturing said host cells; and performing nucleic acid sequence analysis of the transcript markers. This method can be used for gene discovery purposes or transcript imaging purposes. This method can be combined with PCR based technology for the rapid nucleic acid sequence analysis of the transcript markers.

In another aspect of the present invention, a method is provided for generating and analyzing non-contiguous transcript markers derived from the 5' most and 3' most end of individual cDNAs, comprising the steps of obtaining a cDNA library comprising individual cDNAs having a first restriction endonuclease site at the 5' most end and at the 3' most end for a restriction endonuclease that digests nucleic acid within an expected distance from the first endonuclease recognition site, and a second endonuclease restriction site; subjecting the cDNAs to digestion with the first endonuclease thereby creating linearized cDNAs containing transcript markers from the 5' most end and the 3' most end of the individual cDNAs; self-ligating said linearized cDNAs thereby joining the transcript markers from the 5' most and 3' most ends to create cDNAs containing non-contiguous transcript markers; transforming said linearized cDNAs in a host cell and culturing said host cells; isolating said cDNA from said host cells; digesting the cDNA with the second restriction endonuclease thereby excising the non-contiguous transcript markers; ligating said excised transcript markers to a second vector; transforming said second vector containing the transcript markers in a host cell and culturing said host cells; and performing nucleic acid sequence analysis of the transcript markers.

In another aspect of the present invention, a method is provided for the rapid, sequence-specific identification of cDNAs derived from a human mRNA population, comprising the steps of obtaining a cDNA library comprising individual cDNAs wherein said cDNAs contain first restriction endonuclease sites for an endonuclease having a 4 base pair recognition site and wherein said cDNAs are cloned into a first vector lacking the first restriction endonuclease sites; subjecting the cDNA library to digestion with the first restriction endonuclease thereby creating linearized cDNAs containing a portion of the original cDNA; ligating an adapter to said linearized cDNAs wherein the adapter contains a second restriction endonuclease site for an endonuclease that cleaves within an expected range of its recognition site thereby creating cDNAs containing two non-contiguous transcript markers joined by the adapter; digesting the cDNAs with the second restriction endonuclease thereby excising the transcript markers from said cDNAs; concatenating said excised transcript markers; ligating said concatenated transcript

markers to a second vector; transforming said second vector containing the transcript markers in a host cell and culturing said host cells; and performing nucleic acid sequence analysis on the transcript markers in said host cells.

The present invention also provides novel methods and vectors used in preparation of cDNA libraries containing transcript markers. The present invention also provides novel, high-throughput methods for obtaining complete nucleotide sequence information on transcript markers. In one embodiment of the present invention, cDNA libraries containing a serial arrangement of multiple transcript markers are constructed and subjected to high-throughput nucleic acid sequence analysis in a multi-well format using polymerase chain reaction (PCR) technology and specialized PCR primers.

In additional aspects of the present invention, the cDNA libraries are constructed to bias the cDNA population toward rare cDNAs. In one aspect of the present invention, a cDNA library is constructed by normalization techniques and in another, the cDNA library is constructed using subtractive hybridization techniques. In yet another aspect of the present invention, the cDNA library has been constructed from mRNA treated with demethylating agents, such as, 5-aza-2' deoxycytidine and 5-azacytidine, which induces the transcription of silent genes. In an additional aspect of the present invention, the cDNA library has been constructed using oligo dT primers and in another aspect, the cDNA library is constructed using random primers.

BRIEF DESCRIPTION OF DRAWINGS

Figure 1 illustrates a general schematic of cDNA library construction.

Figure 2 illustrates Array of Transcript Markers (ATM) Strategy I for constructing a cDNA library containing an array of transcript markers derived from the 5' most end of a cDNA (5' transcript markers).

Figure 3 illustrates ATM Strategy II for constructing a cDNA library containing an array of 5' transcript markers.

Figures 4A-4B illustrate ATM Strategy III for constructing a cDNA library containing an array of 5' transcript markers.

Figures 5A-5B illustrate ATM Strategy IV for constructing a cDNA library containing an array of 5' transcript markers.

Figures 6A-6B illustrate ATM Strategy V for constructing a cDNA library containing an array of 5' transcript markers.

Figure 7 illustrates a strategy for simultaneously obtaining two non-contiguous transcript

markers from both the 5' and 3' end of cDNA.

Figure 8 illustrates a strategy for obtaining non-contiguous transcript markers from random areas of a cDNA.

Figure 9 illustrates a sequence specific approach to the identification of all transcribed genes from an mRNA population.

Figure 10 is a list of 12 known sequences identified by the method illustrated in Figure 2.

Figure 11 illustrates the toxicity curve for THP1 cells treated for 3 days with 5-aza-2'-cytidine.

Figures 12A-12B illustrate examples of Type IIs restriction endonucleases useful for generating ATM libraries. BpmI, BsgI and Eco57I are three restriction endonuclease that cleave 16 bp away from their recognition site with a 2 nucleotide 3' overhang. 'BpmI is illustrated in Figure 12A. N's represent nucleotides found adjacent to the restriction site (e.g., the 5' end of a cDNA). Treatment with T4 DNA polymerase removes the 3' extension which leaves 14 bp derived from the adjacent sequence. Figure 12B illustrates that the restriction endonuclease BSMFI cleaves 14 bp away from its recognition site with a 4 nucleotide 5' overhang. This end may be filled in by treatment with a DNA polymerase which results in 14 bp derived from adjacent sequence.

Figure 13 illustrates a vector useful in the strategy of Figure 8 which has all restriction endonuclease sites having a 4 base pair recognition site removed.

Figures 14A-14B illustrate a typical concatenated array of transcript markers.

Figure 15 illustrates an abundance profile of THP cells treated with 5-aza-2' deoxycytidine. The black-shaded area represents control clones, the gray shaded area represents 5-aza-2' deoxycytidine treated cells and the gray line represents the abundance profile of the 5-aza-2' deoxycytidine treated cells.

MODES FOR CARRYING OUT THE INVENTION

Before the present compounds, variants, formulations and methods for making and using such are described, it is to be understood that this invention is not limited to the particular compounds, variants, formulations or methods described, as such variants, formulations and methodologies may, of course, vary. It is to be understood that the terminology used herein is for the purpose of describing particular embodiments only, and it is not intended to be limiting since the scope of the present invention will be limited only by the appended claims.

It must be noted that as used in the specification and the appended claims, the singular

forms "a", "an" and "the" include plural references unless the context clearly dictates otherwise which will be known to those skilled in the art or will become known to them upon reading this specification.

I. Definitions

As used herein, the term "transcript marker derived from a cDNA library" refers to an isolated polynucleotide derived from an individual cDNA and being preferably from about 10 base pairs to about 20 base pairs in length derived from an individual cDNA.

As used herein, the term "non-contiguous transcript marker from an individual cDNA" refers to two polynucleotides which are not adjacent to one another under naturally occurring conditions, but which are constructed to exist in tandem, with each polypeptide being preferably from about 10 base pairs to about 20 base pairs in length.

As used herein, the term "3' transcript marker" or "transcript marker from the 3' most end of a cDNA" refers to an isolated polynucleotide derived from the 3' most end of an individual cDNA and being preferably from about 10 base pairs to about 20 base pairs in length.

As used herein the term "5' transcript marker" or "5' transcript marker derived from a cDNA library" refers to an isolated 5' most nucleic acid sequence of a cDNA which is preferably about 10 base pairs to about 20 base pairs in length. As used herein, "5' most" means that a 5' transcript marker may represent the 5' end of the full-length coding region of a cDNA and may include 5' untranslated sequences. Alternatively, a 5' most transcript marker may represent an internal coding region of a individual cDNA. Each 5' transcript marker can reflect the expression of an individual cDNA.

As used herein, the term "adapter" refers to a synthetic fragment of nucleic acid which is ligated to a cDNA and which may contain recognition sites for restriction endonucleases.

As used herein the term "5' adapter" refers to a synthetic fragment of nucleic acid which is ligated onto the 5' end of cDNAs prior to ligation to a vector. In the present invention, the 5' adapter contains a first restriction endonuclease site which digests nucleic acid at a expected distance from its enzymatic recognition site and at least one other restriction endonuclease site. The second restriction endonuclease site is for a restriction endonuclease which digests the cDNA to form blunt ends or 5' or 3' overhangs. In a preferred embodiment of the present invention, digestion of the cDNA with the first and second restriction enzymes excises transcript markers which are at least 20 base pairs in length and contain at least 14 base pairs of cDNA sequence and 6 base pairs of adapter sequence.

As used herein the term "array(s) of transcript markers" or "ATM" refers to the collection or serial arrangement of transcript markers prepared by concatenation of individual transcript markers. As used herein, an "ATM cDNA library" is a cDNA library containing transcript markers as inserts. The inserts may be individual inserts, multiple inserts or a serial arrangement of multiple inserts which may be in sense or antisense orientation.

As used herein the term "full-length" coding region refers to the cDNA sequence for the entire transcribed mRNA for a particular protein from initiating methionine to the poly A tail.

As used herein the term "type IIs restriction endonuclease" or "type IIs enzyme " refers to that category of restriction endonucleases that digests nucleic acid at an expected distance from the enzymatic recognition site. Preferred examples of type IIs enzymes for use in the present invention are those which digest nucleic acid at least 10 base pairs from the recognition site or which digest nucleic acid less than 10 base pairs from the recognition site but can be filled in enzymatically to give a 10 base pair transcript marker. Examples of such Type IIs enzymes include, but are not limited to BpmI, BsgI Eco57I and BsmFI. Examples of type IIs restriction endonucleases is illustrated in Figures 12A-12B. As will be understood by those of skill in the art, it is possible to alter enzymatic digestion conditions to change the nucleic acid digestion position of the type IIs restriction endonuclease.

As used herein the term "type IIsg restriction endonuclease" or "type IIsg enzyme" refers to that category of restriction endonucleases that digests nucleic acid within an expected range of its enzymatic recognition site. Examples of type IIsg restriction enzymes include, but are not limited to BcgI.

As used herein the term "concatenating" refers to the process of ligating multiple transcript markers prior to ligation in a cloning vector.

As used herein the term "normalized cDNA library" or "normalized library" refers to a cDNA library constructed in such a manner as to reduce the redundancy in high- level abundance cDNAs.

As used herein the term "subtractive hybridization" (when referring to construction of a cDNA library) refers to a process wherein a first population of nucleic acid is hybridized with a second labelled population of nucleic acid (driver) and the resultant nucleic acid hybrids removed to completion thereby identifying and isolating a set of nucleic acid sequences unique to the first population of nucleic acid which may be used in the construction of a cDNA library.

As used herein the term "selected set" of random primers refers to a set of primers

designed to anneal to a specific region of mRNA. In a preferred embodiment herein, the selected set of random primers is designed to anneal to the 5' most region of mRNA used as starting material for cDNA synthesis.

As used herein the term "transcript imaging" refers to a method of determining the relative abundance of individual transcript markers in a cDNA library. The term relative abundance refers to the number of times an individual transcript marker appears relative to the total number of transcript markers identified.

As used herein the term "non-amplified growth" of host cells refers to growth conditions which allow for uniform growth of recombinant host cells.

As used herein, the term "high abundance" or "high-level abundance" messages exist at greater than 10,000 copies per cell.

As used herein, the term "mid abundance" or "mid-level abundance" species exist at 100 to 400 copies per cell.

As used herein, the term "low abundance" or "low-level abundance" species are found at less than 15 copies per cell. This low-abundance class represents 20 to 50 percent of the unique transcripts in the cell.

As used herein the term "wobble primer" refers to a sequencing primer degenerate at the 3'-end to allow sequencing for all three possible bases following the poly A tail.

The present invention relates to rapid, high-throughput gene discovery methods that facilitate genome closure and to methods for analyzing gene expression patterns. The present invention is based, in part, upon the discovery of methods for the generation of transcript markers derived from the 5' end of cDNAs contained within a cDNA library. The present invention is also based, in part, upon the discovery of methods for the generation of two discontinuous transcript markers derived from both the 5' and 3' ends of cDNAs contained within a cDNA library.

The transcript markers of the present invention are derived from previously constructed cDNA libraries. The nucleotide information contained within transcript markers can be used to identify novel transcripts or to provide the basis for the design of PCR primers useful in extending and identifying a transcript marker nucleotide sequence contained within a cDNA library. CDNA libraries of the present invention can be constructed by methods which equalize the population of cDNAs or bias the population of cDNAs toward rare cDNAs, e.g. by normalization techniques, subtraction techniques, treatment with demethylating agents and

treatment with differentiating agents, for example.

The present invention also provides novel methods for achieving genome closure which combine rapid nucleic acid sequence identification of transcript markers from a cDNA with subsequent extension of the sequence of the identified transcript markers using PCR technology. Methods of the present invention may also be used to provide a transcript image of specific tissues or biological samples.

II. Construction of cDNA libraries

CDNA libraries for use in the methods of the present invention may be prepared by methods described in Maniatis et al. (1982, Molecular Cloning: a Laboratory Manual, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York) or any means known to those of skill in the art.

CDNA libraries may be constructed to bias the population of individual cDNAs toward desired coding regions. A cDNA library constructed with oligo dT primers would be expected to contain nucleic acid sequences from the 3' coding region of a cDNA, as well as 3' untranslated regions. Additionally, a cDNA library constructed with oligo dT primers may also contain part or all of the entire coding region of a particular cDNA. A cDNA library constructed with random primers would be expected to contain nucleic acid sequences from all parts of the coding region of a cDNA, including the 5' most nucleic acid sequence of the full-length coding region.

CDNA libraries constructed with selected sets of random primers, such as primers which specifically prime first strand cDNA synthesis toward the 5' end, such as with the incorporation of CapFinder™ PCR construction kit (Clontech K1051-1), are desirable for obtaining 5' transcript markers from the 5' most end of a putative transcript or cDNA.

For gene discovery purposes, preferred methods for the construction of cDNA libraries include the use of random primers or a selected set of random primers designed to prime cDNA synthesis from the 5' end of the mRNA; the use of cell lines treated with demethylating compounds, such as, 5-aza-2'-deoxycytidine, as starting material for the preparation of mRNA; the use of cell lines treated with compounds which induce differentiation, such as retinoic acid; normalization or equalization methods; the use of subtractive hybridization techniques in the construction of cDNA libraries; and any method that induces the transcription of silent genes or allows for the identification of rare cDNAs, such as size fractionation of specific cDNA populations expected to contain rare cDNAs.

Preferred methods for the construction of cDNA libraries intended for transcript imaging

purposes are those which produce an unbiased population of cDNAs and would include a step for the non-amplified or uniform growth of host cells used in constructing the cDNA library. Such growth conditions are described in Current Protocols in Molecular Biology "Amplification of Cosmid and Plasmid Libraries" Unit 5.10. (1987).

In the normalization process, the prevalence of high-abundance cDNA clones decreases dramatically, clones with mid-level abundance are relatively unaffected, and clones for rare transcripts are effectively increased in abundance. In the Soares et al. (1994, supra) normalization procedure, the abundance levels of individual cDNA clones have been equalized by a kinetic re-annealing hybridization. This approach is designed to reduce the initial 10,000-fold variation in individual cDNA frequencies in order to achieve abundances within one order of magnitude while maintaining the overall sequence complexity of the library.

CDNA libraries prepared by normalization techniques are not an accurate reflection of the source tissue's gene-expression profile; however cDNA libraries produced by normalization techniques may provide a source of low abundance transcripts and therefore, would be useful for gene discovery purposes.

A variety of normalization techniques are known by those of skill in the art for the production of normalized cDNA libraries. Weissmann S.M. (1987 Mol. Bio. Med. 4:133-143) describe a method based on hybridization to genomic DNA wherein the frequency of each hybridized cDNA in the resulting normalized library would be proportional to that of each corresponding gene in the genomic cDNA. Ko. (1990, Nucleic Acid Res. 18:5705-5711) and Patanjali et al (1991, Proc. Natl. Acad. Sci. 88:1943-1947) describe a kinetic approach to constructing cDNA libraries. Soares (WO 95/08647, filed September 23, 1994 and published March 30, 1995) describe a method to normalize a 3' directional cDNA library.

Subtractive hybridization of nucleic acids is a method to isolate the coding sequences of a gene which are differentially expressed such as during development or in disease states. CDNA libraries prepared by subtractive hybridization techniques are described in Rubenstein et al (1990, Nucleic Acids Research, 18:4833-4842); Travis et al. (1988, Proc. Natl. Acad. Sci. USA, 85:1696-1700).

Obtaining cDNA libraries from cells treated with 5-aza-2'-deoxycytidine should enhance the discovery of rare genes and genes expressed in specialized cell types from which it is difficult to isolate and/or prepare DNA. A non-toxic concentration of 5-aza-2'-deoxycytidine can be pre-determined through titration/toxicity assays as illustrated in Figure 11. 5-aza-2'-deoxycytidine

has been shown to induce the transcription of silent genes through demethylation (Hsieh et al *supra*). The preferred amount of 5-aza-2'-deoxycytidine to use is that amount that induces the transcription of silent genes without being toxic to the cell. This method can be coupled to subtractive methods to enhance further the discovery of novel transcripts or genes.

II. Construction of ATM Libraries

ATM libraries containing transcript markers can be constructed using any of the methods described in Figures 2-8. Figure 2 illustrates ATM Strategy I for constructing a cDNA library containing an array of transcript markers derived from the 5' most end of a cDNA (5' transcript markers). In Strategy I, a cDNA library is constructed using an adapter containing a BpmI site and a PvuII restriction site. The constructed cDNA library is digested with BpmI and PvuII to isolate 20 bp 5' transcript markers (14 base pairs of cDNA and 6 base pairs from the adapter) from the cDNA library, the isolated markers are treated with T4 DNA polymerase to yield blunt ends and the markers are concatenated to form an array or serial arrangement of multiple markers. The concatenated array of 5' transcript markers is then used to create a library containing an array of 5' transcript markers. In a variation of this method, individual transcript markers can be subjected to subtractive hybridization methods to bias the population toward rare transcript markers.

Another strategy is shown in Figure 3 which illustrates ATM Strategy II for constructing a cDNA library containing an array of 5' transcript markers. In Strategy II, a cDNA library is constructed using an adapter containing a BpmI site and a PvuII restriction site. The constructed cDNA library is digested with BpmI and a second degenerate adapter containing a PvuII site is ligated onto the BpmI site. After PCR amplification of the template and PvuII digestion, 25 bp transcript markers are isolated. The isolated transcript markers are concatenated to form arrays, ligated into a vector and transformed into host cells to create a library containing an array of 5' transcript markers.

ATM Strategy III, as shown in Figures 4A-4B illustrates the construction of a cDNA library containing an array of 5' transcript markers. In Strategy III, a cDNA library is constructed using an adapter containing a BpmI site and a PvuII restriction site. The constructed cDNA library is digested with BpmI and a second degenerate adapter containing a PvuII site is ligated onto the BpmI site. Sense RNA is transcribed from the template and the mixture is subjected to DNAse treatment. First and second strand cDNA is synthesized from the template. The double stranded cDNA is digested with PvuII which yields 25 bp transcript markers. The transcript

markers are concatenated to form arrays, the arrays are cloned into a vector and the vector transformed into host cells to create a library containing an array of 5' transcript markers.

ATM strategy IV, as shown in Figures 5A-5B, illustrates ATM Strategy IV for constructing a cDNA library containing an array of 5' transcript markers. In strategy IV, a cDNA library is constructed using an adapter containing a BpmI and PvuII restriction site. The constructed cDNA library is digested with BpmI and PvuII to isolate 20 bp 5' transcript markers from the cDNA library, the isolated markers are treated with T4 DNA polymerase to yield blunt ends and the markers are concatenated to form an array or serial arrangement of multiple markers. Adapters are added on to the ends of the transcript marker arrays and the arrays are amplified using PCR technology. The arrays are cloned into a vector to create a library containing arrays of the transcript markers (ATM).

Another strategy is shown in Figures 6A-6B which illustrate ATM Strategy V for constructing a cDNA library containing an array of 5' transcript markers. In strategy V, a cDNA library is constructed using an adapter containing a BpmI and PvuII restriction site. The constructed cDNA library is digested with BpmI and PvuII to isolate 20 bp 5' transcript markers from the cDNA library, the isolated markers are treated with T4 DNA polymerase to yield blunt ends and the markers are concatenated to form an array or serial arrangement of multiple markers. At this point, the arrays are ligated into a plasmid vector, subjected to PCR to amplify the transcript markers, and cloned into a vector to create a cDNA library containing PCR amplified arrays of 5' transcript markers.

Figure 7 illustrates a strategy for simultaneously obtaining two non-contiguous transcript markers from both the 5' and 3' end. In this strategy, first strand cDNA synthesis is performed using a modified random hexamer. The hexamer is designed to provide directionality. Second strand cDNA synthesis is performed by standard means and an adapter containing a BpmI and PvuII site is ligated onto both ends of the cDNA. The cDNA is ligated to a vector modified to delete all BpmI restriction sites. The vector containing the cDNA is transformed into a host cell to create a cDNA library and plasmid DNA is isolated and treated with BpmI, thereby creating individual linearized cDNAs containing both a 5' and 3' transcript marker. The linearized cDNA is treated with T4 polymerase to blunt end, self-ligated and re-transformed to create a library. The plasmid DNA of the library is isolated and digested with PvuII to excise the nucleic acid containing non-contiguous transcript markers from both the 5' and 3' end. The excised transcript markers are concatenated and cloned to create a library containing a serial arrangement of non-

contiguous transcript markers from the 5' and 3' end which can be subjected to high-throughput nucleic acid sequence analysis. In another variation of this method, the vector is constructed to contain 5' and 3' BpmI sites at a cloning site thereby providing a means to excise the transcript marker from the cDNA.

Figure 8 illustrates a strategy for obtaining non-contiguous transcript markers from random areas of a cDNA. In this strategy, mRNA is converted to cDNA following standard procedures. The cDNA is ligated to a vector that has been constructed to have restriction endonuclease sites for restriction endonucleases having 4 base pair recognition sites removed. The cDNA library is amplified and the plasmids prepared. The cDNA library is digested with a restriction endonuclease having a 4 base pair recognition site and the digested plasmid is purified. A 12 base pair adapter containing a BcgI site, which digests nucleic acid within a range of its recognition site, is ligated onto the linearized cDNA, and the cDNA is transformed into a host cell. The cDNA library is digested with BcgI resulting in the release of a 36 base pair fragment from each cDNA which originally contained the 4 base pair restriction endonuclease site. The fragments may be ligated into a vector directly or concatenated and ligated into a vector and subjected to sequence specific analysis.

The ATM libraries of the present invention may contain a single transcript marker per individual clone or multiple transcript markers constructed in a serial arrangement. As illustrated in the figures, transcript markers may be concatenated, PCR amplified and then ligated into a vector or concatenated, ligated into a vector and then PCR amplified.

In a preferred embodiment of the present invention, transcript markers that are excised from the cDNA library contain a cDNA portion and a synthetic adapter portion. In a preferred embodiment disclosed herein and as shown in Figure 2, the cDNA portion is at least 14 base pairs in length and the adapter portion is 6 base pairs which are designed to be asymmetric. The adapter portion provides the means for determining the sense orientation of the transcript marker in the vector as well as a means for determining the beginning of each transcript marker excised. Nucleic acid sequencing of the excised transcript markers can be performed to create a nucleic acid data set. For nucleic acid sequence analysis purposes, the nucleic acid adapter portion of the transcript marker can be subtracted or removed from the transcript marker nucleic acid data set.

The ATM markers may exist in sense or antisense orientation in the vector. The presence of an nucleic acid fragment which provides directionality, such as an asymmetric adapter portion, provides the means for discerning the sense from the anti-sense stand and provides the means for

determining the beginning of each transcript marker. Figure 14A illustrates sequence data from a single clone containing an array of 6 transcript markers. The 6 bp "spacer" DNA sequences that distinguish one transcript marker from another are underlined. The spacer sequence "CTGGAG" indicates that the immediately adjacent 14 bp to the right is a transcript marker (sense strand). The spacer sequence "CTCCAG" indicates that adjacent 14 bp to the left is a transcript marker (antisense strand). Figure 14B provides a list of the six 14 bp transcript markers (sense strand) without the spacer DNA sequence derived from the array in (14a). Asterisks (*) indicate sequences which are the reverse complement of the sequence actually found in the array.

IV. Method for Nucleic Acid Sequencing

Nucleic acid sequencing of transcript markers can be performed by any means known to those of skill in the art. Methods for cDNA sequencing employ such enzymes as the Klenow fragment of cDNA polymerase I, Sequenase® (US Biochemical Corp, Cleveland OH)), Taq polymerase (Perkin Elmer, Norwalk CT), thermostable T7 polymerase (Amersham, Chicago IL), or combinations of recombinant polymerases and proofreading exonucleases such as the ELONGASE Amplification System marketed by Gibco BRL (Gaithersburg MD). Preferably, the process is automated with machines such as the Hamilton Micro Lab 2200 (Hamilton, Reno NV), Peltier Thermal Cycler (PTC200; MJ Research, Watertown MA) and the ABI 377 cDNA sequencers (Perkin Elmer).

V. PCR Methods

Numerous PCR methods are known to those of skill in the art that would facilitate isolation, amplification and/or extension of nucleic acid sequences in the 5' or 3' direction. Gobinda et al (1993; PCR Methods Applic 2:318-22) disclose "restriction-site" polymerase chain reaction (PCR) as a direct method which uses universal primers to retrieve unknown sequence adjacent to a known locus. First, cDNA is amplified in the presence of primer to a linker sequence and a primer specific to the known region. The amplified sequences are subjected to a second round of PCR with the same linker primer and another specific primer internal to the first one. Products of each round of PCR are transcribed with an appropriate RNA polymerase and sequenced using reverse transcriptase.

Inverse PCR can be used to amplify or extend sequences using divergent primers based on a known region (Triglia T et al (1988) Nucleic Acids Res 16:8186). Adapters are ligated onto cDNAs which then allow cDNAs to be circularized. The intramolecular ligation products then serve as PCR templates. The method uses several restriction enzymes to generate a suitable

fragment in the known region of a gene. The fragment is then circularized by intramolecular ligation and used as a PCR template.

Capture PCR (Lagerstrom M et al (1991) PCR Methods Applic 1:111-19) is another method which may be used. The method involves PCR amplification of cDNA fragments adjacent to a known sequence in human and yeast artificial chromosome cDNA. Capture PCR also requires multiple restriction enzyme digestions and ligations to place an engineered double-stranded sequence into an unknown portion of the cDNA molecule before PCR. Another PCR method which may be used to retrieve sequences is that of Parker JD et al (1991; Nucleic Acids Res 19:3055-60).

Capillary electrophoresis may be used to analyze the size or confirm the nucleotide sequence of sequencing or PCR products. Systems for rapid sequencing are available from Perkin Elmer, Beckman Instruments (Fullerton CA), and other companies. Capillary sequencing may employ flowable polymers for electrophoretic separation, four different fluorescent dyes (one for each nucleotide) which are laser activated, and detection of the emitted wavelengths by a charge coupled devise camera. Output/light intensity is converted to electrical signal using appropriate software (eg. Genotyper™ and Sequence Navigator™ from Perkin Elmer) and the entire process from loading of samples to computer analysis and electronic data display is computer controlled. Capillary electrophoresis is particularly suited to the sequencing of small pieces of cDNA which might be present in limited amounts in a particular sample. The reproducible sequencing of up to 350 bp of M13 phage cDNA in 30 min has been reported (Ruiz-Martinez MC et al (1993) Anal Chem 65:2851-2858).

VI. Method for determining Genome Closure

The present invention provides novel methods for achieving genome closure by combining rapid nucleic acid sequence identification of transcript markers from a cDNA with subsequent extension of the sequence of the identified transcript markers using PCR technology.

Transcript markers constructed by the method illustrated in Figure 8 which have been concatenated and ligated to a vector provide for the sequence specific identification of 15-20 transcript markers per vector. In one embodiment described herein, PCR technology is used to extend the sequence of a transcript marker in an outward direction as described in US Patent Application 08/487,112, filed June 7, 1995, specifically incorporated by reference. One primer is synthesized to initiate extension in the antisense direction (XLR) and the other is synthesized to extend sequence in the sense direction (XLF). Primers allow the extension of the known

sequence "outward" generating amplified nucleotide sequences containing new, unknown nucleotide sequence for the region of interest. As shown below, PCR primers which contain the sequence of a known adapter joining the non-contiguous transcript markers and all possible combinations of nucleotides for the 5 nucleotide positions flanking the adapter are used in PCR reactions to extend the sequence of the identified transcript markers using PCR technology.

N N N N N 16 base pair adapter N N N N N

5' _____ 3'

Considering that for each N position, there could be 4 possible nucleotides, a total of 1024 individual primers would be required for each 5' and 3' extension for a total of 1,048,576 combinations which will be specific enough to amplify any DNA.

CDNA libraries which have been prepared with oligo dT primers will allow for the identification of the 3' most part of the cDNA and part of the coding region or the complete 5' most end of a cDNA. CDNA libraries constructed with random primers constructed in combination with techniques that result in a high representation of the 5' ends of cDNA, such as the Cap-Finder™ PCR construction kit (Clontech), will allow for the amplification of the 5' ends of genes.

In another method illustrated in Figure 9, genome closure can be achieved by the identification of all transcribed genes from an mRNA source. As illustrated in Figure 9, step 1, a cDNA library containing a biotinylated poly A tail and having a Not I restriction site is constructed by standard means and cloned into a vector from which restriction endonuclease sites for 4 base pair restriction endonucleases have been removed. Multiple cDNA libraries derived from a variety of mRNA sources can be pooled to provide one sample. In Figure 9, step 2, the cDNA is digested with a restriction endonuclease which has a 4 base pair recognition site and the digested cDNA is captured by streptavidin-beads (Figure 9 step 3). The cDNA is digested with Not I which removes the biotin and ligated into a vector having a Type IIs restriction endonuclease site for MboII (Figure 9, step 4). Other Type IIs restriction endonucleases, such as BpmI, BsgI Eco57I and BsmFI, can be used. The cDNA is digested with MboII which cuts 8 base pairs into the cDNA. A known linker of sufficient size for PCR amplification is cloned into the Mbo II restriction site and the cDNA pools are subjected to PCR analysis using the primers described above. Generation of the PCR-extension products can either be sequenced directly or with a wobble primer, ie, a sequencing primer degenerate at the 3'-end to allow for all 3 possible bases following the poly A tail, or religated and cloned into a vector and then subjected to nucleic

acid sequencing.

VII. Transcript Imaging

Transcript imaging is a method for evaluating changes in gene expression caused by factors such as disease progression, pharmacologic treatment and aging. Transcript imaging is accomplished by sequencing several thousand clones from a particular tissue or cell type and electronically recording the abundance levels for each mRNA species identified. Electronic manipulations can then be done to examine which mRNAs are up- or down-regulated, or unchanged.

Transcript imaging can be achieved by using transcript markers produced by the present invention. After nucleic acid sequencing of the transcript markers is accomplished, the abundance levels for each transcript marker are electronically recorded. Electronic manipulations can then be performed to examine which transcript markers are up- or down- regulated, or unchanged.

INDUSTRIAL APPLICABILITY

I. Construction of cDNA Libraries

The following example describes the construction of a cDNA library. The first step is to isolate mRNA from a desired biological tissue or cell source. The mRNA is then used in the synthesis of cDNA.

RNA Isolation

RNA is isolated using guanidinium isothiocyanate and 2-mercaptoethanol lysis, followed by ultracentrifugation over a cesium chloride gradient to obtain total RNA (Chirgwin et al). Alternatively, total RNA can be isolated using acid/phenol extraction (Chowzisky et al) and polyadenylated RNA can be isolated directly using a biotinylated oligo dT primer. An optical density measurement is taken to assess the quantity of total RNA isolated, and an aliquot is run on an electrophoresis gel to assess the quality and integrity of the RNA. The RNA is then stored until needed at -80°C, which prevents degradation.

In order to obtain cleaner total RNA, each sample is treated with DNAse and acid phenol, followed by precipitation and washing. The RNA is again run on an electrophoresis gel to make sure it is free of genomic DNA contamination. Subsequent selection of polyadenylated (polyA) RNA is done with either an oligo(dT)-based affinity column or Oligotex™ latex microspheres. The quality of the isolated mRNA is checked confirmed, and the sample is used in cDNA library construction.

cDNA Library Construction

Synthesis of first-strand cDNA is initiated using a poly(dT) primer that is complementary to the polyA stretch at the 3' end of most transcripts; and the reverse transcriptase enzyme. The primer used in this reaction contains a restriction enzyme recognition site (NotI) that permits directional insertion into an appropriate cloning vector. Second-strand cDNA synthesis is based on the method developed by Gubler and Hoffman (1983). RNAse H nicks the RNA/cDNA hybrid created in the reverse transcription reaction, creating priming sites for E. coli DNA polymerase to synthesize second-strand cDNA. The gaps in the second strand are ligated together using E. coli DNA ligase.

After the ends of the cDNA are blunted with T4 or Pfu DNA polymerase, an adaptor is ligated to the double-stranded cDNA. This oligonucleotide, which contains an EcoRI-compatible sticky end, allows for directional cloning on the cDNA once digestion is complete with NotI, the restriction enzyme site found at the 3' terminus of the cDNA. The cDNA is then size-fractionated to remove very short cDNAs, which would inhibit the generation of highly complex libraries. At this point, the cDNA is ligated into a plasmid vector system and transformed into bacterial cells for propagation.

II. Construction of ATM cDNA libraries

Three cDNA libraries were constructed containing an adapter having the Type IIs restriction endonuclease, Bpm I, at the 5' end of each cDNA insert.

Two ug each of poly A' RNA from human colon, human prostate, and a single species control RNA (bacterial chloramphenicol transferase) were reverse transcribed using an oligo-dT primer and Superscript reverse transcriptase according to manufacturer's instructions (Gibco Superscript Plasmid System). Following second strand synthesis, 3 ug of the adapter containing the Bpm I site were ligated to each cDNA.

The adapter containing the Bpm I site was prepared previously in the following manner: two oligonucleotides (5' AATTCAGCTGGAG and 5' phos-CTCCAGCTG) were synthesized and purified by HPLC and polyacrylamide gel electrophoresis (New England Biolabs). Equimolar amounts of the two oligos were combined in annealing buffer (20 mM Tris, pH 7.4, 2 mM $MgCl_2$, 50 mM NaCl), boiled, and allowed to cool to <30 degrees C in a heating block. Following adapter ligation, the cDNA was digested with Not I, fractionated over a Sepharose CL-4B column and cloned into a pSPORT vector (LTI, Inc.).

To create an ATM library we followed the procedure as outlined in ATM Strategy 1, as

illustrated in Figure 2, starting with prostate mRNA containing the Bpm I adapter described above. This library was transformed into E. coli strain DH10B by electroporation and allowed to grow either on LB plates or LB broth each supplemented with carbenicillin. Transformed cells were collected either by scraping (plates) or by centrifugation (media) and plasmid DNA was isolated (Promega or Qiagen systems). Greater than 100 ug of plasmid DNA was then digested with Bpm I and Pvu II. Multiple aliquots of 20 ug of plasmid were digested for 1.5 hours using 14 U of Bpm I (New England Biolabs) in a 150 ul volume at 37 degrees C. Thirty units of Pvu II (New England Biolabs) were then added and digestion proceeded for an additional hour at 37 degrees C. The digested DNA was fractionated on 20% TBE acrylamide. A 20 bp band was excised, the DNA was recovered by electroelution or by incubating the crushed gel slices in tris/EDTA overnight. The transcript markers were treated with T4 DNA polymerase and cloned either directly into pSPORT or concatenated.

2. **Analysis of ATM libraries**

ATM libraries were made according to ATM Strategy 1, illustrated in Figure 2. Vector background of these libraries was determined by screening randomly selected clones from the respective libraries by restriction enzyme digestion of plasmids with Bpm I and PvuII, thereby releasing transcript marker inserts (n= number of clones screened). Library size is determined by subtracting the background from the number of individual colonies or transformants that are generated from a single ligation reaction of the transcript markers to pSport vector. Average number of markers per clone as shown in Table I is determined by both insert size and DNA sequence verification.

TABLE I

library	vector background	library size	avg.#markers/clone
Prostate 1	10%(n=10)	1.1×10^6	3.8
Prostate 2	0% (n=12)	0.95×10^6	4.9
Prostate 3	0% (n=12)	0.12×10^6	3.8

Arrays of transcript markers from the prostate ATM library were PCR amplified for 30 cycles using M13 forward and reverse primers. These arrays were size selected on a 6% acrylamide gel and cloned into a pSPORT-derived vector. Twenty-seven random clones were sequenced from a total of 131 total transcript markers analyzed. There was an average of 4.9 markers per clone with a range of 4-7 markers/clone. Table II illustrates the average marker size per number of clones.

TABLE II

marker size	number (%)
12	2 (1.5)
13	6 (4.6)
14	102 (77.9)
15	14 (10.7)
16	1 (0.8)
14/15	6 (4.6)

The marker size "14/15" refers to an ambiguous situation where there are 29 bp (instead of 28 bp) with markers which are concatenated back to back as shown:

spacer |-----29 bp--two markers back to back------| spacer

CTGGAGNNNNNNNNNNNNNNNNNNNNNNNNNNNCTCCAG

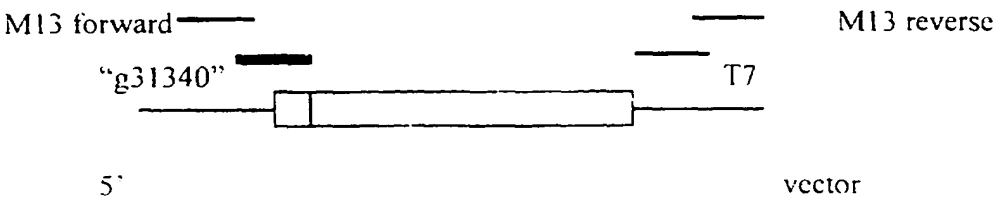
Because the frequency of 13 and 16 bp markers is much lower than 14 and 15 bp markers, the 29 bp between the spacer sequences are assumed to be composed of one 14 bp and one 15 bp marker, rather than one 13 bp and one 16 bp marker. Thus, the true total number of 14 bp markers was 105 (102+3; 80.2%) and the true total number of 15 bp markers was 17 (14+3; 13%).

III. Isolation of a full length cDNA

A transcript marker (5' ccgagagtcgtcgg) was identified from a prostate ATM library which corresponds to the gene apoferritin H (GenBank GI numbers: g31340, g31342, g28434). Based on the published sequence, this marker is present at the 5' end of the mRNA, 14 nucleotides downstream from the start of transcription and 181 nucleotides upstream of the start of the coding sequence. The entire mRNA is predicted to be about 0.9 kb.

As illustrated below, to isolate the apoferritin H cDNA from the ATM library, a gene specific primer ("g31340") that contains 7 nucleotides of the Bpm I adapter and the 14 nucleotides of the transcript marker (5' gctggagccgagagtcgtcgg) was designed. The cDNA inserts from 1 ug of the prostate ATM library were amplified by 30 cycles of PCR using the M13 forward and reverse primers. This PCR reaction was diluted 1:50 in water and 1 ul was

reamplified for 33 cycles using the nested T7 and "g31340" primers. A 0.9 kb band was isolated, gel purified and cloned. DNA sequencing confirmed the identity of the cloned gene as full length apoferritin H. Alternatively, the 0.9 kb PCR product is sequenced directly with appropriate primers.



IV. High Throughput Isolation of cDNA clones

High throughput isolation of cDNA clones from an ATM library is achieved in the following manner. First, a master pool of insert cDNA is created from an ATM library by PCR amplification using primers found in the vector (e.g., M13 forward and reverse). Second, gene specific primers are synthesized in 96 well arrays (Gibco). Third, aliquots of gene specific primers, PCR reagents, and the master cDNA pool are aliquoted to 96 PCR wells for PCR and subsequently analyzed by gel electrophoresis. In addition, an initial screen for successful PCR reactions is accomplished by doing real time flourescent detection of PCR products. With this technique only those reactions which give a significant fluorescent signal above background would then be analyzed by gel electrophoresis.

V. Extension of Transcript Markers to Full Length

The nucleic acid sequence of transcript markers can be used to design oligonucleotide primers for extending a partial nucleotide sequence to full length or for obtaining 5' sequences from genomic libraries. One primer is synthesized to initiate extension in the antisense direction (XLR) and the other is synthesized to extend sequence in the sense direction (XLF). Primers allow the extension of the known sequence "outward" generating amplified nucleotide sequences containing new, unknown nucleotide sequence for the region of interest (US Patent Application 08/487,112, filed June 7, 1995, specifically incorporated by reference). The initial primers are designed from the cDNA using OLIO® 4.06 Primer Analysis Software (National Biosciences), or another appropriate program, to be 22-30 nucleotides in length, to have a GC content of 50% or more, and to anneal to the target sequence at temperatures about 68°-72° C. Any stretch of nucleotides which would result in hairpin structures and primer-primer dimerizations is avoided.

The original, selected cDNA libraries, or a human genomic library, is used to extend the

sequence. A genomic library is most useful to obtain 5' upstream regions. If more extension is necessary or desired, additional sets of primers are designed to further extend the known region.

By following the instructions for the XL-PCR kit (Perkin Elmer) and thoroughly mixing the enzyme and reaction mix, high fidelity amplification is obtained. Beginning with 40 pmol of each primer and the recommended concentrations of all other components of the kit, PCR is performed using the Peltier Thermal Cycler (PTC200; MJ Research, Watertown MA) and the following parameters:

Step 1	94° C for 1 min (initial denaturation)
Step 2	65° C for 1 min
Step 3	68° C for 6 min
Step 4	94° C for 15 sec
Step 5	65° C for 1 min
Step 6	68° C for 7 min
Step 7	Repeat steps 4-6 for 15 additional cycles
Step 8	94° C for 15 sec
Step 9	65° C for 1 min
Step 10	68° C for 7:15 min
Step 11	Repeat step 8-10 for 12 cycles
Step 12	72° C for 8 min
Step 13	4° C (and holding)

A 5-10 μl aliquot of the reaction mixture is analyzed by electrophoresis on a low concentration (about 0.6-0.8%) agarose mini-gel to determine which reactions were successful in extending the sequence. Bands thought to contain the largest products were selected and cut out of the gel. Further purification involves using a commercial gel extraction method such as QIAQuick™ gel extraction (QIAGEN Inc). After recovery of the DNA, Klenow enzyme was used to trim single-stranded, nucleotide overhangs creating blunt ends which facilitate religation and cloning.

After ethanol precipitation, the products are redissolved in 13 μl of ligation buffer, 1μl T4-DNA ligase (15 units) and 1μl T4 polynucleotide kinase are added, and the mixture is incubated at room temperature for 2-3 hours or overnight at 16° C. Competent E. coli cells (in 40 μl of appropriate media) are transformed with 3 μl of ligation mixture and cultured in 80 μl of SOC medium (Sambrook J et al, supra). After incubation for one hour at 37° C, the whole transformation mixture is plated on Luria Bertani (LB)-agar (Sambrook J et al, supra) containing 2xCarb. The following day, several colonies are randomly picked from each plate and cultured in 150 μl of liquid LB/2xCarb medium placed in an individual well of an appropriate, commercially-available, sterile 96-well microtiter plate. The following day, 5 μl of each

overnight culture is transferred into a non-sterile 96-well plate and after dilution 1:10 with water. 5 μl of each sample is transferred into a PCR array.

For PCR amplification, 18 μl of concentrated PCR reaction mix (3.3x) containing 4 units of rTth DNA polymerase, a vector primer and one or both of the gene specific primers used for the extension reaction are added to each well. Amplification is performed using the following conditions:

Step 1	94° C for 60 sec
Step 2	94° C for 20 sec
Step 3	55° C for 30 sec
Step 4	72° C for 90 sec
Step 5	Repeat steps 2-4 for an additional 29 cycles
Step 6	72° C for 180 sec
Step 7	4° C (and holding)

Aliquots of the PCR reactions are run on agarose gels together with molecular weight markers. The sizes of the PCR products are compared to the original partial cDNAs, and appropriate clones are selected, ligated into plasmid and sequenced.

VI. 5 - aza-2' Deoxycytidine Treatment of Cells

5-aza-2'-deoxycytidine induces transcription of silent genes, presumably by demethylating cytosines in CpG islands which are regulatory regions located upstream of most genes. btaining libraries from cells treated with 5-aza-2'-deoxycytidine will enhance the discovery of rare genes and genes expressed in specialized cell types difficult to isolate and prepare RNA from.

Methods

THP1 cells at a density of 1.1 million cells per ml were treated for three days with 0.8 micromolar 5-aza-2'-deoxycytidine. The medium used for growth conditions was Iscove's modified DMEM with 10% Fetal Bovine Serum.

HNT precursor cells at 80% confluency were treated for three days with 0.35 micromolar 5-aza-2'-deoxycytidine. The medium used for growth conditions was Iscove's modified DMEM.

Because 5-aza-2'-deoxycytidine has been shown to be toxic in cultured cells and animal, initial experiments were conducted to assess the toxicity of 5-aza-2'-deoxycytidine on hNT and THP1 cells to establish conditions where the cells would survive and RNA could be recovered. Around 1 micro molar 5-aza-2'-deoxycytidine is a concentration typically used to induce silent gene transcription. A concentration of 0.8 micro molar was selected for THP1 cells and 0.35 micro molar for hNT cells.

Results

Northern analysis to measure the RNA levels of two identified genes was performed on the cells to verify that 5-aza-2'-deoxycytidine was inducing gene transcription under the conditions used. Significant induction of both identified genes by 5-aza-2'-deoxycytidine was obtained in the hNT cells. Many genes were found to be induced by 5-aza-2'-deoxycytidine in both THP1 cells and hNT cells.

The following 3 genes were in the cDNA library constructed from mRNA from THP cells treated with 5-aza-2'deoxycytidine library and not in the control cDNA library: enBank g29382, human BBC1 mRNA; GenBank g337507, human ribosomal protein S25 mRNA; and g184553 human insulinoma pig-analog mRNA.

The following genes were in the cDNA library constructed from mRNA from THP cells treated with 5-aza-2'deoxycytidine library and in the control cDNA library and found upregulated in THP1: GenBank g182055 human neutrophil elastase mRNA, 3' end; g36143 human mRNA for ribosomal protein S11; g793842 human mRNA for ribosomal protein L29; g28976 human mRNA for azurocidin; and g348911 human glycoprotein mRNA.

The following 3 genes were in the cDNA library constructed from mRNA from hNT cells treated with 5-aza-2'deoxycytidine library and not in the control cDNA library: g190233 human acidic ribosomal phosphoprotein P1; g436217 human mRNA (KIAA0037) for ORF; and g385936 hinge=OXPHOS system complex III.

An additional result of 5-aza-2'-deoxycytidine treatment was a decrease in the expression of many genes, particularly more abundant mRNAs.

VII. Construction of the Plasmid pIIEZ-1

Plasmid pIIEZ-1 was derived from pUC19 vector by removing the 991 base pairs from the restriction endonuclease sites SspI through AfI3. A 90 base pair synthetic polylinker containing type IIs restriction endonuclease sites were ligated to the remaining pUC19 vector, as shown in Figure 13.

The type IIs polylinker was made by annealing the two synthetic oligomers P-II-1 and P-II-2. The oligomer sequences are:

P-II-1

5'

CATGTGCGGCCGCGGCGCGCCGTCAGCTGCACAGATGCGAGCTCCAGGCATTCATCC

TTAAGTACTTCAGTAGACGTCCCTGCAGGTGAATTC

P-II-2

5'

GAATTCACCTGCAGGGACGTCTACTGAAGTACTTAAGGATGAATGCCTGGAGCTCGC

ATCTGTGCAGCTGACGGCGCGCCGCGGCCGCA

After annealing, the double stranded linker has a 5' blunt end which is ligated into the SspI restriction endonuclease site and a 4 bp (5' CATG) extension which is complimentary to the AfI3 restriction endonuclease site.

As illustrated in Figure 13, the polylinker contains 4 type IIs restriction endonuclease sites which can be used for generation of 14 bp transcript markers.

Standard PCR primer directed site specific mutagenesis is carried out to eliminate all type IIs restriction endonuclease sites and all 4 base pair restriction endonuclease sites.

VIII. Genome Closure

A method directed toward genome closure involves the systematic identification of all transcribed genes. The method involves the generation of cDNA's with a defined 5'-end. generation of defined priming sites for extension PCR amplification from within the cloned cDNAs, systematic amplification of all cDNAs, and sequencing of all extension products.

Generation of cDNA's with a defined 5'-end

As illustrated below, after conversion of mRNA into cDNA with standard methods, the cDNA's are digested with a 4 base pair cutter. The cDNAs are ligated into the vector displaying a type IIs restriction endonuclease site at one end.

1) Generation of cDNAs

*Biotin

----------cDNA----------TTTTTTTTTTTTTT NotI

----------AAAAAAAAAAAAAA

2) Digestion with Hpa II

*Biotin

------V-----cDNA--------V---------TTTTTTTTTTTTTT NotI

--------V----------V-------AAAAAAAAAAAAAA

3) Capture of cDNA's with Streptavidin-beads

*Biotin

CGG----------TTTTTTTTTTTTTT NotI Y

C----------AAAAAAAAAAAAAA bead

4) Ligation of adapter/linker

adapter/linker

*Biotin

TCGACTAGTACGAAGA CGG----------TTTTTTTTTTTTTT NotI Y

GATCATGCTTCTGC C----------AAAAAAAAAAAAAA bead

*Biotin

TCGACTAGTACGAAGACGG----------TTTTTTTTTTTTTT NotI Y

GATCATGCTTCTGCC----------AAAAAAAAAAAAAA bead

5) Not I digest, ligation into pUC 18 vector (Not I/Sal I)

----, ----V----------TTTTT NotI

| |

---------- vector----------

6) Digestion with Bbs I, cuts 6bp into cDNA

Bbs I (digest with Bbs I)

GAAGACGC NNNNNNNNN

|||||||| |||||

CTTCTGCGNNNN NNNNN

| |

| |

| |

SalI----------vector----------NotI

7) Bbs I (fill in with Klenow)

```
      GAAGACGCNNNN                    NNNNNNNNN
      ||||||||||||                    ||||||||
5     CTTCTGCGNNNN                    NNNNNNNNN
        |                                       |
        |                                       |
        |                                       |
        SalI----------vector----------NotI
```

10

Generation of defined priming sites for extension PCR amplification from within the cloned cDNAs.

The cDNAs are cloned into a vector, amplified and digested with the type IIs restriction

15 endonuclease. Linkers are cloned into the linearized plasmid/cDNA fragment.

8)Bbs I CATGATCATGG

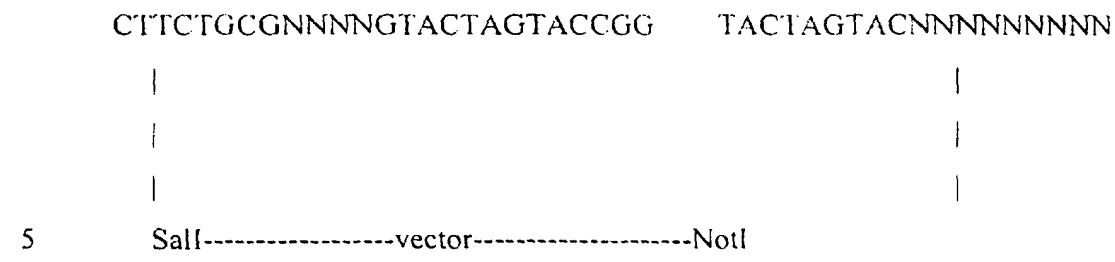
```
        ||||||||||
        GTACTAGTACCGG
```

20 9)Ligate on linkers

```
      GAAGACGCNNNNCATGATCATGG    GGCCATGATCATGNNNNNNNNN
      ||||||||||||||||||||||     ||||||||||||||||||||
      CTTCTGCGNNNNGTACTAGTACCGG  GGTACTAGTACNNNNNNNNN
        |                                       |
25      |                                       |
        |                                       |
        SalI----------vector----------NotI
```

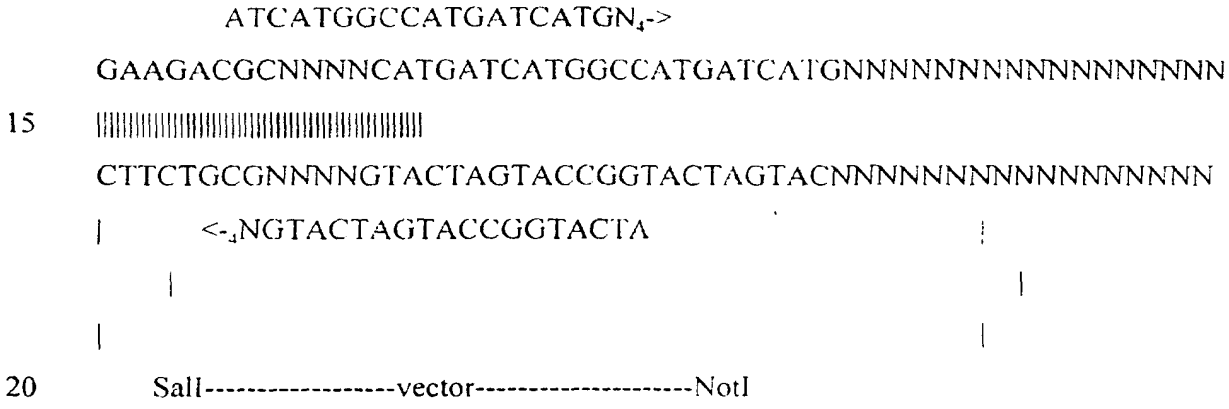
30 10)Treat with T4-DNA polymerase in presence of (dA, dT)

```
      a) GAAGACGCNNNNCATGATCAT    GGCCATGATCATGNNNNNNNNN
      ||||||||||||||||||||||     ||||||||||||||||||||
```



Systematic amplification of all cDNAs

Extension PCR as described in Example V with combinations of defined primer sets (up to 10-fold degeneracy can be achieved, depending on the type IIs enzyme selected in the step 4 adapter). Therefore selective amplification of every cDNA present in the library can be achienved in a systematic manner.



Since NNNN and NNNN are symmetrical combinations for primer pairs are given (e.g. 5' NNNN 3' = 5' AGCA 3' then

5'ATCATGGCCATGATCATGAGCA 3' will form a pair with
5'ATCATGGCCATGATCATGTCGT 3'

Any PCR-reaction will cover 2 primer combinations.

combination 1:
5'ATCATGGCCATGATCATGAGCA 3'
5'ATCATGGCCATGATCATGTCGT 3'

and

combination 2:

5'ATCATGGCCATGATCATGAGCA 3'

5'ATCATGGCCATGATCATGTCGT 3'

Example of selective amplification with a 4 fold degenerate primer set resulting in 128 pools of cDNAs

To cover all possible combinations of 4 bases, 256 oligonucleotides will be synthesized (5'ATCATGGCCATGATCATGNNNN 3'). Since every PCR-reaction covers 2 combinations, 128 reactions will cover all possible combinations. A range of 200-300 PCR-products/reaction will be expected for a specific tissue. Since this is an inverse PCR approach, the products are sequenced directly with the wobble primer, increasing the resolution by a factor 3. The PCR-products then are directly thermo-cycled with the wobble primers resulting in a higher resolution. One colour is choosen for each nucleotide with no terminator included in the reaction. The resulting products are run on a sequencing gel.

All publications and patents mentioned in the above specification are herein incorporated by reference. Various modifications of the described method and system of the invention will be apparent to those skilled in the art without departing from the scope and spirit of the invention. Although the invention has been described in connection with specific preferred embodiments, it should be understood that the invention as claimed should not be unduly limited to such specific embodiments. Indeed, various modifications of the above-described modes for carrying out the invention which are obvious to those skilled in the field of molecular biology or related fields are intended to be within the scope of the following claims.

SEQUENCE LISTING

(1) GENERAL INFORMATION

(i) APPLICANT: INCYTE PHARMACEUTICALS, INC.

(ii) TITLE OF THE INVENTION: METHODS FOR GENERATING AND ANALYZING TRANSCRIPT MARKERS

(iii) NUMBER OF SEQUENCES: 50

(iv) CORRESPONDENCE ADDRESS:
(A) ADDRESSEE: Incyte Pharmaceuticals, Inc.
(B) STREET: 3174 Porter Drive
(C) CITY: Palo Alto
(D) STATE: CA
(E) COUNTRY: U.S.
(F) ZIP: 94304

(v) COMPUTER READABLE FORM:
(A) MEDIUM TYPE: Diskette
(B) COMPUTER: IBM Compatible
(C) OPERATING SYSTEM: DOS
(D) SOFTWARE: FastSEQ Version 1.5

(vi) CURRENT APPLICATION DATA:
(A) PCT APPLICATION NUMBER: To Be Assigned
(B) FILING DATE: Filed Herewith

(vii) PRIOR APPLICATION DATA:
(A) APPLICATION NUMBER: US 08/723,646
(B) FILING DATE: 03-OCT-1996

(viii) ATTORNEY/AGENT INFORMATION:
(A) NAME: Billings, Lucy J.
(B) REGISTRATION NUMBER: 36,749
(C) REFERENCE/DOCKET NUMBER: IN-0001 PCT

(ix) TELECOMMUNICATION INFORMATION:
(A) TELEPHONE: 650-855-0555
(B) TELEFAX: 650-845-4166

(2) INFORMATION FOR SEQ ID NO:1:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 13 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:1:

AATTCAGCTG GAG 13

(2) INFORMATION FOR SEQ ID NO:2:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 9 base pairs
(B) TYPE: nucleic acid

(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:2:

CTCCAGCTG 9

(2) INFORMATION FOR SEQ ID NO:3:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 41 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:3:

CTGGAGNNNN NNNNNNNNNN NNNNNNNNNN NNNNNCTCCA G 41

(2) INFORMATION FOR SEQ ID NO:4:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 14 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(vii) IMMEDIATE SOURCE:
(A) LIBRARY: GenBank
(B) CLONE: 31340, 31342, 28434

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:4:

CCGAGAGTCG TCGG 14

(2) INFORMATION FOR SEQ ID NO:5:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 21 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(vii) IMMEDIATE SOURCE:
(A) LIBRARY: GenBank
(B) CLONE: 31340

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:5:

GCTGGAGCCG AGAGTCGTCG G 21

(2) INFORMATION FOR SEQ ID NO:6:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 13 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:6:

AATTCAGCTG GAG 13

(2) INFORMATION FOR SEQ ID NO:7:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 35 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:7:

AATTCAGCTG GAGNNNNNNN NNNNNNNNNC AGCTG 35

(2) INFORMATION FOR SEQ ID NO:8:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 31 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:8:

GTCGACCTCN NNNNNNNNNN NNNNNGTCGA C 31

(2) INFORMATION FOR SEQ ID NO:9:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 16 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:9:

CCCGAGGTCG ACTTAA 16

(2) INFORMATION FOR SEQ ID NO:10:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 22 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(vi) ORIGINAL SOURCE:

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:10:

CTGGAGNNNN NNNNNNNNNN NN 22

(2) INFORMATION FOR SEQ ID NO:11:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 20 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:11:

GACCTCNNNN NNNNNNNNNN 20

(2) INFORMATION FOR SEQ ID NO:12:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 21 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:12:

GTTGAATACT CATACTCTTC C 21

(2) INFORMATION FOR SEQ ID NO:13:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 20 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:13:

GCTGGCCTTT TGCTCACATG 20

(2) INFORMATION FOR SEQ ID NO:14:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 60 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:14:

GAATTCACCT GCAGGGACGT CTACTGAAGT ACTTAAGGAT GAATGCCTGG AGCTCGCATC 60

(2) INFORMATION FOR SEQ ID NO:15:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 29 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:15:

TGTGCAGCTG ACGGCGCGCC GCGGCCGCA 29

(2) INFORMATION FOR SEQ ID NO:16:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 120 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:16:

CAACCACCCG GGCCCTCCAG CTGGAGGAAA AAATGCTAGG CTGGAGGGCT GATGTTTTCC 60
CTGGAGCTAG TTCTAGATCG CTGGAGCTGC GCCCGGCCCG GGGACCGGCT ATCCCTCCAG 120

(2) INFORMATION FOR SEQ ID NO:17:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 14 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:17:

GGCCCGGGTG GTTG 14

(2) INFORMATION FOR SEQ ID NO:18:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 14 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:18:

GAAAAAATGC TAGG 14

(2) INFORMATION FOR SEQ ID NO:19:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 14 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:19:

GGCTGATGTT TTCC 14

(2) INFORMATION FOR SEQ ID NO:20:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 14 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:20:

CTAGTTCTAG ATCG 14

(2) INFORMATION FOR SEQ ID NO:21:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 14 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:21:

CTGCGCCCGG CCCG 14

(2) INFORMATION FOR SEQ ID NO:22:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 14 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:22:

GGATAGCCGG TCCC 14

(2) INFORMATION FOR SEQ ID NO:23:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 93 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:23:

CATGTGCGGC CGCGGCGCGC CGTCAGCTGC ACAGATGCGA GGTCCAGGCA TTCATCCTTA 60
AGTACTTCAG TAGACGTCCC TGCAGGTGAA TTC 93

(2) INFORMATION FOR SEQ ID NO:24:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 89 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:24:

GAATTCACCT GCAGGGACGT CTACTGAAGT ACTTAAGGAT GAATGCCTGG AGCTCGCATC 60
TGTGCAGCTG ACGGCGCGCC GCGGCCGCA 89

(2) INFORMATION FOR SEQ ID NO:25:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 14 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:25:

TTTTTTTTTT TTTT 14

(2) INFORMATION FOR SEQ ID NO:26:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 14 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:26:

AAAAAAAAAA AAAA 14

(2) INFORMATION FOR SEQ ID NO:27:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 16 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:27:

TCGACTAGTA CGAAGA 16

(2) INFORMATION FOR SEQ ID NO:28:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 14 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:28:

CGTCTTCGTA CTAG 14

(2) INFORMATION FOR SEQ ID NO:29:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 12 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:29:

GAAGACGCNN NN 12

(2) INFORMATION FOR SEQ ID NO:30:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 12 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:30:

NNNNGCGTCT TC 12

(2) INFORMATION FOR SEQ ID NO:31:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 11 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:31:

CATGATCATG G 11

(2) INFORMATION FOR SEQ ID NO:32:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 13 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:32:

GGCCATGATC ATG 13

(2) INFORMATION FOR SEQ ID NO:33:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 23 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:33:

GAAGACGCNN NNCATGATCA TGG 23

(2) INFORMATION FOR SEQ ID NO:34:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 25 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:34:

GGCCATGATC ATGNNNNGCG TCTTC 25

(2) INFORMATION FOR SEQ ID NO:35:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 22 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:35:

GGCCATGATC ATGNNNNNNN NN 22

(2) INFORMATION FOR SEQ ID NO:36:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 20 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:36:

NNNNNNNNNC ATGATCATGG 20

(2) INFORMATION FOR SEQ ID NO:37:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 21 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:37:

GAAGACGCNN NNCATGATCA T 21

(2) INFORMATION FOR SEQ ID NO:38:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 25 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:38:

GGCCATGATC ATGNNNNGCG TCTTC 25

(2) INFORMATION FOR SEQ ID NO:39:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 22 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:39:

GGCCATGATC ATGNNNNNNN NN 22

(2) INFORMATION FOR SEQ ID NO:40:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 18 base pairs
(B) TYPE: nucleic acid

(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:40:

NNNNNNNNNC ATGATCAT 18

(2) INFORMATION FOR SEQ ID NO:41:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 22 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:41:

ATCATGGCCA TGATCATGNN NN 22

(2) INFORMATION FOR SEQ ID NO:42:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 51 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:42:

GAAGACGCNN NNCATGATCA TGGCCATGAT CATGNNNNNN NNNNNNNNNN N 51

(2) INFORMATION FOR SEQ ID NO:43:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 51 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:43:

NNNNNNNNNN NNNNNNNCAT GATCATGGCC ATGATCATGN NNNGCGTCTT C 51

(2) INFORMATION FOR SEQ ID NO:44:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 22 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:44:

ATCATGGCCA TGATCATGNN NN 22

(2) INFORMATION FOR SEQ ID NO:45:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 22 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:45:

ATCATGGCCA TGATCATGAG CA 22

(2) INFORMATION FOR SEQ ID NO:46:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 22 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:46:

ATCATGGCCA TGATCATGTC GT 22

(2) INFORMATION FOR SEQ ID NO:47:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 22 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:47:

ATCATGGCCA TGATCATGAG CA 22

(2) INFORMATION FOR SEQ ID NO:48:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 22 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:48:

ATCATGGCCA TGATCATGTC GT 22

(2) INFORMATION FOR SEQ ID NO:49:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 22 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:49:

ATCATGGCCA TGATCATGAG CA 22

(2) INFORMATION FOR SEQ ID NO:50:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 22 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:50:

ATCATGGCCA TGATCATGTC GT 22

Claims:

1. A method for generating and analyzing transcript markers from the 5' most end of individual cDNAs of a cDNA library comprising the steps of:

a) obtaining a cDNA library comprising individual cDNAs having a first restriction endonuclease site for a restriction endonuclease that digests the cDNA at the 5' most end within an expected distance from its recognition site and a second endonuclease restriction site;

b) subjecting the cDNAs to digestion with the first and the second restriction endonuclease thereby excising transcript markers from the 5' most end of the individual cDNAs;

c) ligating said transcript markers to a vector;

d) transforming said vector containing the transcript markers in a host cell and culturing said host cells; and

e) performing nucleic acid sequence analysis of the transcript markers.

2. The method of Claim 1 further comprising the step of creating blunt ends on the 5' transcript markers after step b.

3. The method of Claim 1 wherein said first restriction enzyme is a Type IIS restriction enzyme.

4. The method of Claim 3 wherein the Type IIS restriction enzyme is selected from the group consisting of BpmI, BsgI, Eco57I and BsmFI.

5. The method of Claim 1 wherein said cDNA library has been constructed by normalization techniques.

6. The method of Claim 1 wherein said cDNA library has been constructed using random primers.

7. The method of Claim 1 wherein said cDNA library has been constructed using olio dT primers.

8. The method of Claim 1 wherein the cDNA library has been constructed from mRNA treated with a demethylating agent.

9. The method of Claim 8 wherein the demethylating agent is 5-aza-2' deoxycytidine.

10. The method of Claim 1 wherein digesting the cDNA library with the second restriction enzyme creates a blunt end.

11. The method of Claim 1 optionally comprising the step of concatenating said 5' transcript markers after step b) thereby forming a serial arrangement of multiple 5' transcript markers.

12. The method of Claim 11 optionally comprising the step of amplifying the serial arrangement of multiple 5' transcript markers by polymerase chain reaction prior to ligating to a vector.

13. A method for determining a representation of gene expression in a cDNA library comprising the steps of:

a) obtaining a cDNA library comprising individual cDNAs having a first restriction endonuclease site for a restriction endonuclease that digests the cDNA at the 5' most end within an expected distance from its recognition site and a second endonuclease restriction site and wherein the cDNA library has been grown under non-amplified growth conditions;

b) generating and isolating 5' transcript markers from the cDNA library wherein each isolated 5' transcript marker identifies an individual cDNA:

c) ligating the isolated 5' transcript markers to a vector and transforming the vector in a host cell;

d) culturing said host cells;

e) performing nucleic acid sequencing on the 5' transcript markers in said transformed host cells to create a nucleic acid data set; and

f) subjecting the data set to analysis to determine the relative abundance of individual 5' transcript markers thereby determining a representation of gene expression in the cDNA library.

14. The method of Claim 13 comprising concatenating the isolated 5' transcript markers after step b) thereby creating a serial arrangement of multiple 5' transcript markers.

15. The method of Claim 13 wherein said cDNA library has been constructed with olio dT primers.

16. The method of Claim 13 wherein after step b) the isolated 5' transcript markers are blunt ended.

17. A method for rapid nucleic-acid sequence analysis of cDNA comprising the steps of:

a) obtaining a cDNA library comprising individual cDNAs having a first

restriction endonuclease site for a restriction endonuclease that digests the cDNA at the 5' most end within an expected distance from its recognition site and a second endonuclease restriction site;

b) subjecting the cDNA library to digestion with the first restriction endonuclease and the second restriction endonuclease to create 5' transcript markers which identify individual cDNAs;

c) isolating said 5' transcript markers;

d) concatenating said 5' transcript markers to create a serial arrangement of multiple transcript markers;

e) ligating said serial arrangement of multiple transcript markers to a vector and transforming said host cell with the vector;

f) performing nucleic acid sequencing analysis on said serial arrangement of 5' transcript markers and identifying transcript markers;

g) preparing a first and a second PCR primer, wherein the first PCR primer is designed from a transcript marker identified in step f) and the second PCR primer is designed from a section of nucleic acid common to all cDNAs in the cDNA library;

h) subjecting the cDNA library to a polymerase chain reaction using the primers of step g) thereby identifying the cDNA designated by a transcript marker; and

i) performing nucleic acid sequencing on the identified DNA.

18. The method of Claim 17 wherein wherein the polymerase chain reaction of step h) is performed in 96 well plates.

19. The method of Claim 17 optionally comprising the step of amplifying the serial arrangement of multiple 5' transcript markers by polymerase chain reaction after step e).

20. The method of Claim 17 further comprising the step of creating blunt ends on the 5' transcript markers after step c.

21. The method of Claim 17 wherein said first restriction enzyme is a Type IIS restriction enzyme.

22. The method of Claim 17 wherein the Type IIS restriction enzyme is selected from the group consisting of BpmI, BsgI, Eco57I and BsmFI

23. The method of Claim 17 wherein said cDNA library has been constructed by normalization techniques.

24. The method of Claim 17 wherein said cDNA library has been constructed using

random primers.

25. The method of Claim 17 wherein said cDNA library has been constructed using olio dT primers.

26. The method of Claim 17 wherein the cDNA library has been constructed from mRNA treated with a demethylating agent.

27. The method of Claim 26 wherein the demethylating agent is 5- aza-2'-deoycytidine.

28. The method of Claim 17 wherein digesting the cDNA library with the second restriction enzyme creates a blunt end.

29. A method for the rapid, sequence-specific identification of cDNAs derived from a human mRNA population, comprising the steps of:

a) obtaining a cDNA library comprising individual cDNAs wherein said cDNAs contain first restriction endonuclease sites for an endonuclease having a 4 base pair recognition site and wherein said cDNAs are cloned into a first vector lacking the first restriction endonuclease sites;

b) subjecting the cDNA library to digestion with the first restriction endonuclease thereby creating linearized cDNAs containing a portion of the original cDNA;

c) ligating an adapter to said linearized cDNAs wherein the adapter contains a second restriction endonuclease site for an endonuclease that cleaves within an expected range of its recognition site thereby creating cDNAs containing two non-contiguous transcript markers joined by the adapter;

d) digesting the cDNAs of step c) with the second restriction endonuclease thereby excising the transcript markers from said cDNAs;

e) concatenating said excised transcript markers;

f) ligating said concatenated transcript markers to a second vector;

g) transforming said second vector containing the transcript markers in a host cell and culturing said host cells; and

h) performing nucleic acid sequence analysis on the transcript markers in said host cells.

30. The method of Claim 29 optionally comprising concatenating the transcript markers after step e) to create a serial arrangement of multiple transcript markers.

31. The method of Claim 29 wherein the cDNA library has been constructed from mRNA treated with a demethylating agent.

32. The method of Claim 29 wherein the demethylating agent is 5-aza-2' deoxycytidine.

33. The method of Claim 29 wherein the cDNA library has been constructed by normalization techniques.

34. The method of Claim 29 wherein the second endonuclease is BcgI.

35. The method of Claim 29 wherein the nucleic acid sequence analysis comprises amplifying transcript markers in an outward manner using 2 specific primers and sequencing directly using a wobble primer.

36. The vector as shown in Figure 13.

37. A method for generating and analyzing non-contiguous transcript markers derived from the 5' most and 3' most end of individual cDNAs, comprising the steps of:

a) obtaining a cDNA library comprising individual cDNAs having a first restriction endonuclease site at the 5' most end and at the 3' most end for a restriction endonuclease that digests nucleic acid within an expected distance from the first endonuclease recognition site, and a second endonuclease restriction site;

b) subjecting the cDNAs to digestion with the first endonuclease thereby creating linearized cDNAs containing transcript markers from the 5' most end and the 3' most end of the individual cDNAs;

c) self-ligating said linearized cDNAs thereby joining the transcript markers from the 5' most and 3' most ends to create cDNAs containing non-contiguous transcript markers;

d) transforming said linearized cDNAs in a host cell and culturing said host cells;

e) isolating said cDNA from said host cells;

f) digesting the cDNA with the second restriction endonuclease thereby excising the non-contiguous transcript markers;

g) ligating said excised transcript markers to a second vector;

f) transforming said second vector containing the transcript markers in a host cell and culturing said host cells; and

h) performing nucleic acid sequence analysis of the transcript markers.

38. The method of Claim 37 further comprising creating blunt ends on the 5' transcript markers after step b.

39. The method of Claim 37 wherein said first restriction enzyme is a Type IIS restriction enzyme.

40. The method of Claim 37 wherein the Type IIS restriction enzyme is selected from the group consisting of BpmI, BsgI, Eco57I and BsmFI.

41. The method of Claim 37 wherein said cDNA library has been constructed by normalization techniques.

42. The method of Claim 37 wherein said cDNA library has been constructed using random primers.

43. The method of Claim 37 wherein said cDNA library has been constructed using olio dT primers.

44. The method of Claim 37 wherein the cDNA library has been constructed from mRNA treated with a demethylating agent.

45. The method of Claim 44 wherein the demethylating agent is 5-aza-2' deoxycytidine.

46. The method of Claim 37 wherein digesting the cDNA library with the second restriction enzyme creates a blunt end.

47. The method of Claim 37 comprising concatenating said 5' transcript markers after step b) thereby forming a serial arrangement of multiple 5' transcript markers.

48. The method of Claim 47 comprising amplifying the serial arrangement of multiple 5' transcript markers by polymerase chain reaction prior to ligating to a vector.

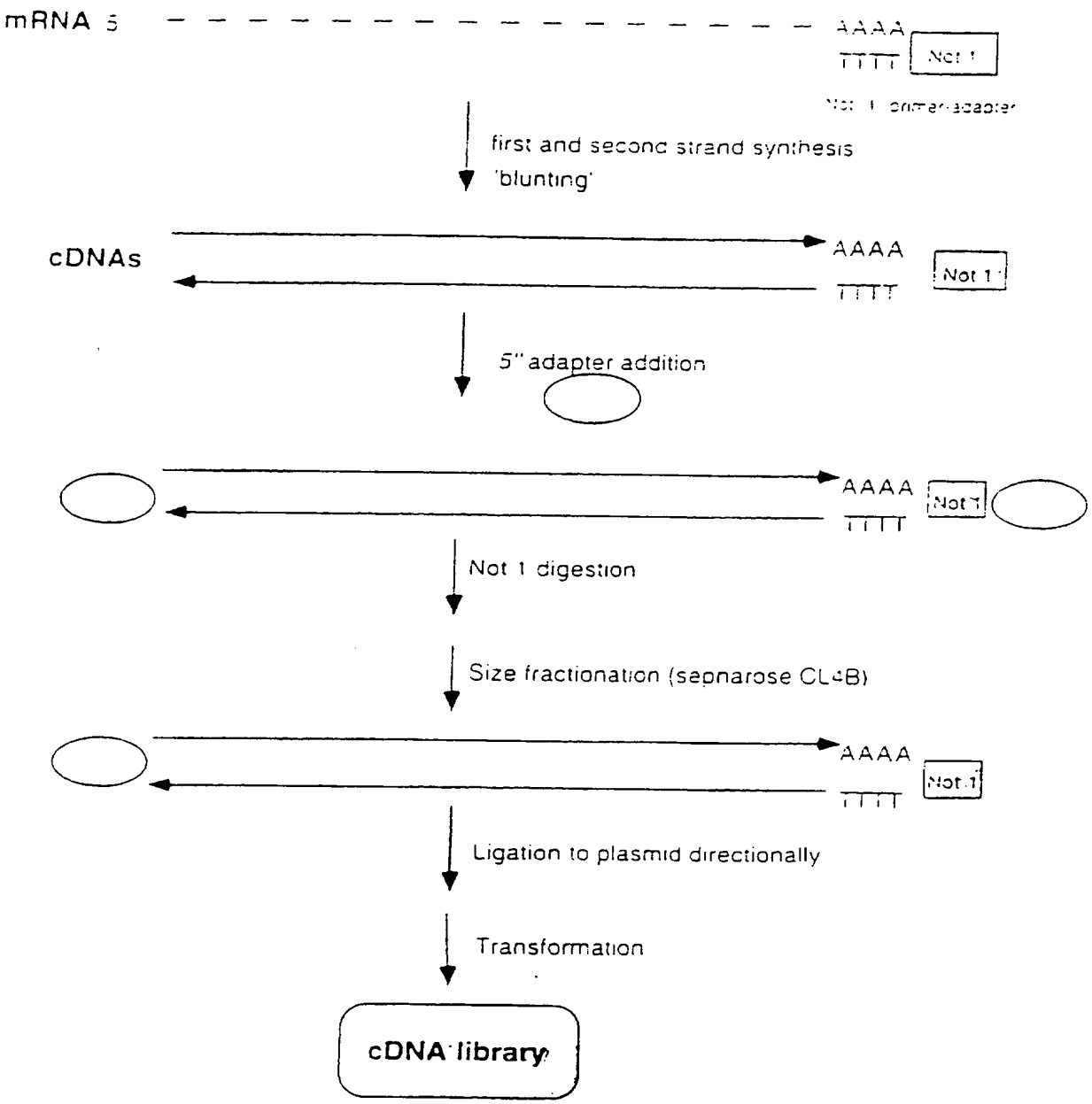
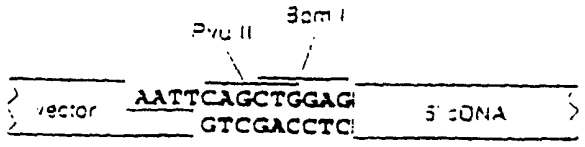

mRNA 5
AAAA
TTTT
Not 1
first and second strand synthesis
'blunting'
cDNAs
AAAA
TTTT
Not 1
5" adapter addition
AAAA
TTTT
Not 1
Not 1 digestion
Size fractionation (sepharose CL4B)
AAAA
TTTT
Not 1
Ligation to plasmid directionally
Transformation
cDNA library

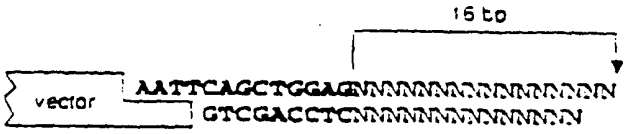
FIGURE 1

1. Construct cDNA library using standard techniques with an EcoRI/BpmI adapter

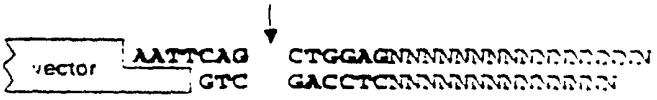


2. Transform library and isolate plasmid DNA

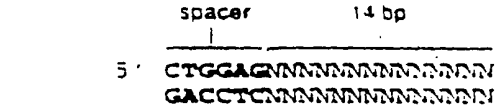
3. Digest DNA with BpmI



4. Digest DNA with PvuII



5. Isolate 20 bp **transcript markers** and treat with T4 DNA polymerase



6. Concatenate blunt **markers** to form **arrays**

7. Clone **arrays** to create **ATM library**

FIGURE 2

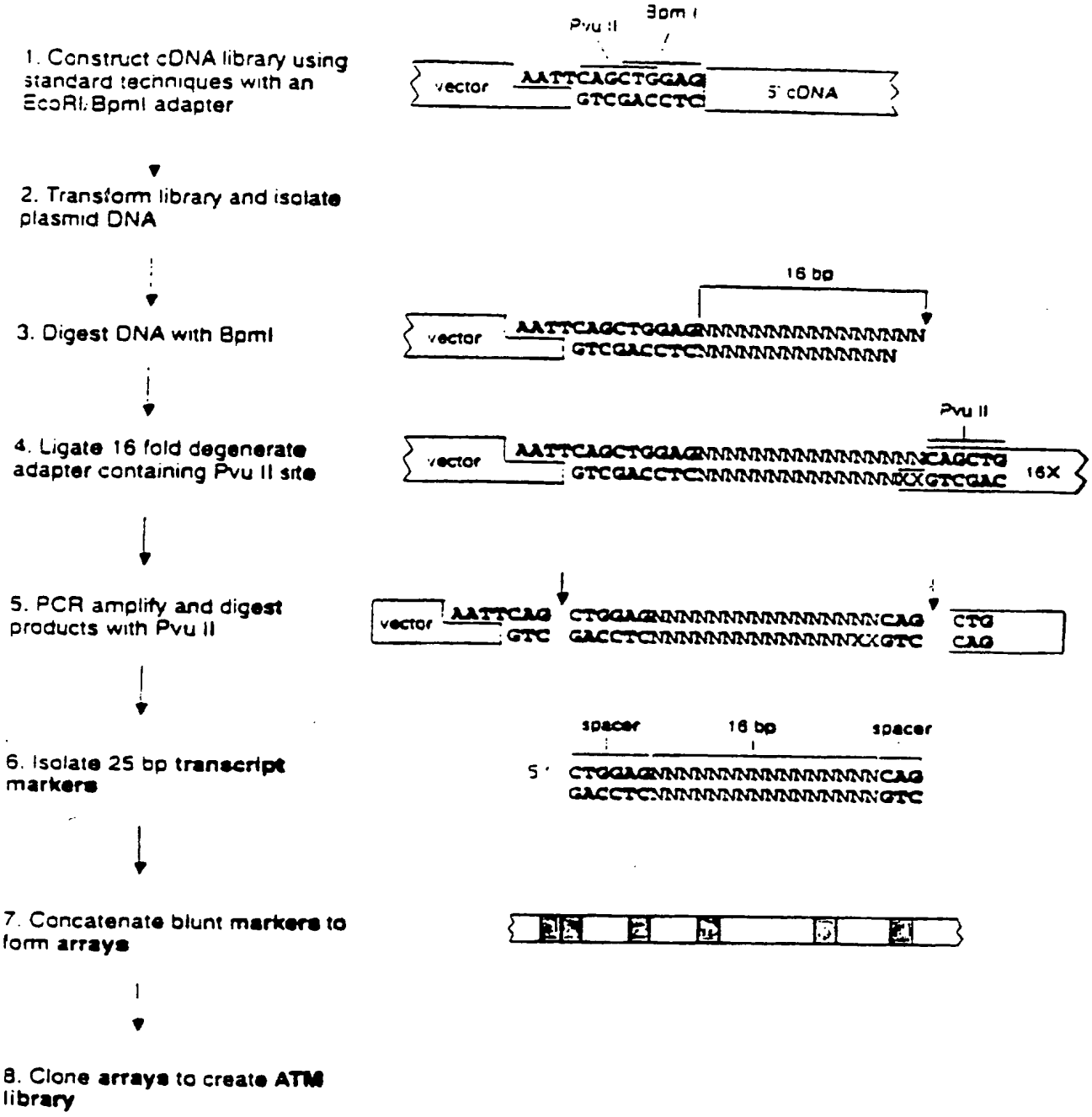


FIGURE 3

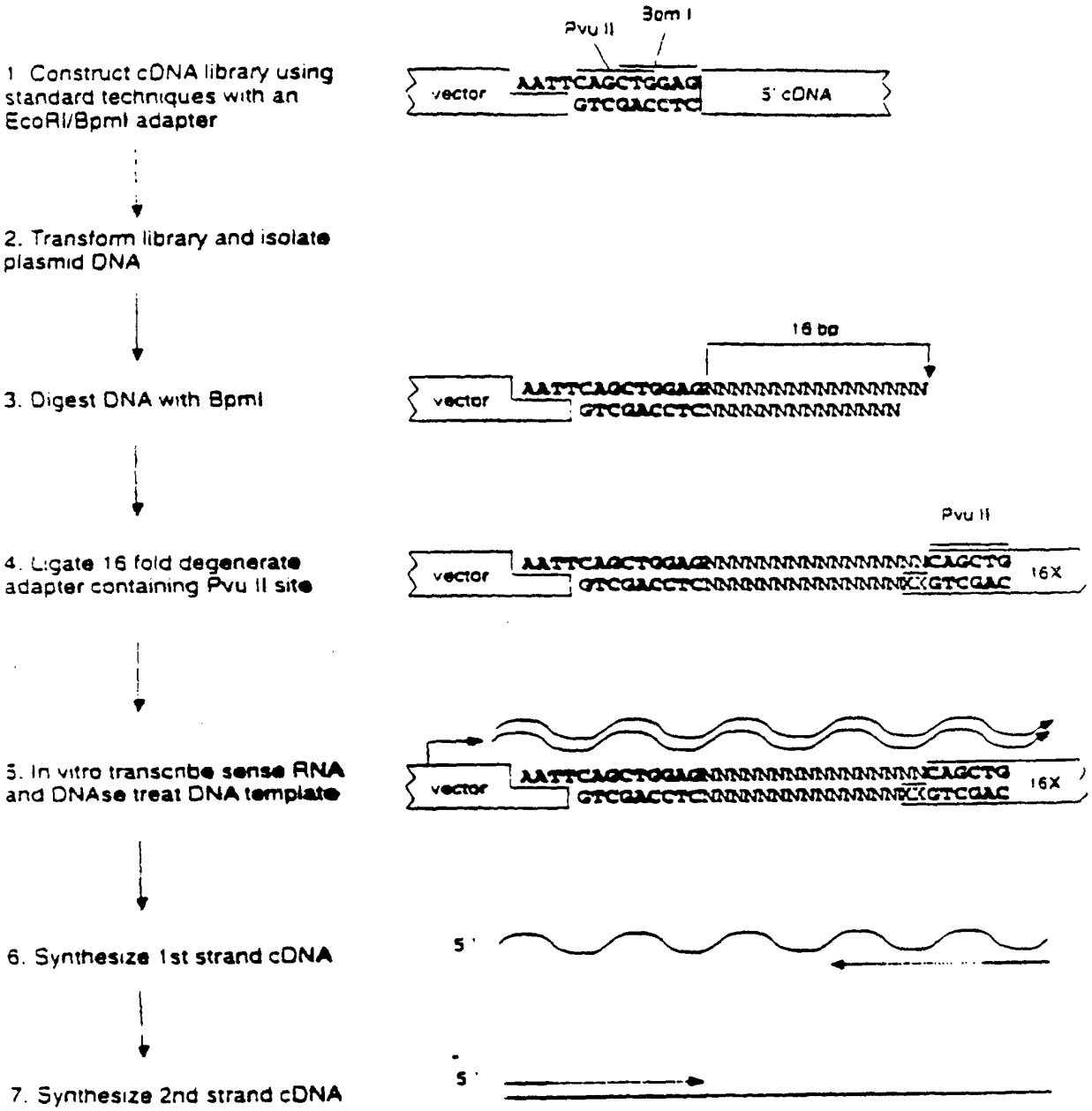

1. Construct cDNA library using standard techniques with an EcoRI/BpmI adapter
Pvu II
BpmI
vector
AATTCAGCTGGAG
GTCGACCTC
5' cDNA
2. Transform library and isolate plasmid DNA
3. Digest DNA with BpmI
16 bp
vector
AATTCAGCTGGAGNNNNNNNNNNNNNNN
GTCGACCTCNNNNNNNNNNNNNN
4. Ligate 16 fold degenerate adapter containing Pvu II site
Pvu II
vector
AATTCAGCTGGAGNNNNNNNNNNNNNNNCAGCTG
GTCGACCTCNNNNNNNNNNNNNNNGTCGAC
16X
5. In vitro transcribe sense RNA and DNAse treat DNA template
vector
AATTCAGCTGGAGNNNNNNNNNNNNNNNCAGCTG
GTCGACCTCNNNNNNNNNNNNNNNGTCGAC
16X
6. Synthesize 1st strand cDNA
5'
7. Synthesize 2nd strand cDNA
5'

FIGURE 4A

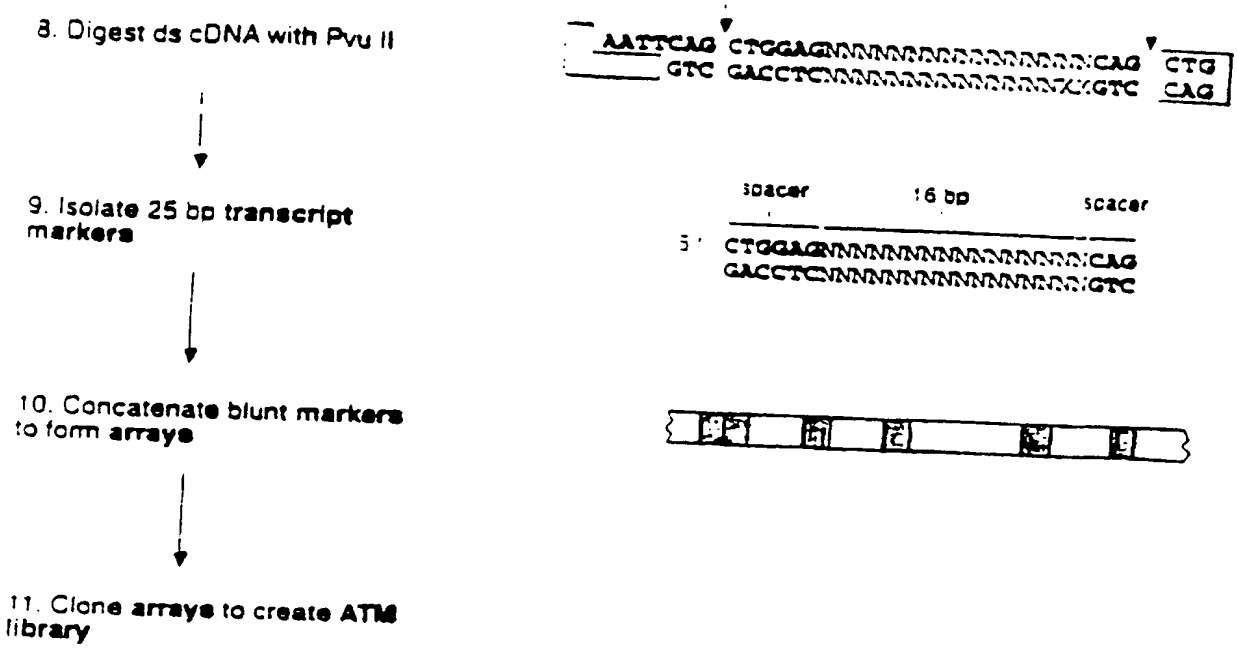
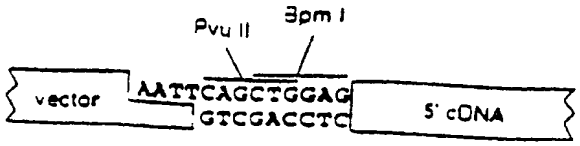

8. Digest ds cDNA with Pvu II
9. Isolate 25 bp transcript markers
10. Concatenate blunt markers to form arrays
11. Clone arrays to create ATM library
AATTCAG CTGGAG
GTC GACCTC
CTG
CAG
spacer
16 bp
spacer
5'
CTGGAG
GACCTC
CAG
GTC

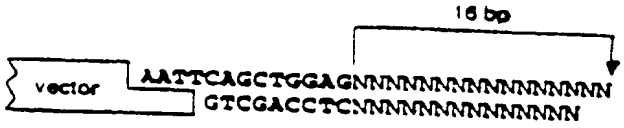
FIGURE 4B

1. Construct cDNA library using standard techniques with an EcoRI/BpmI adapter

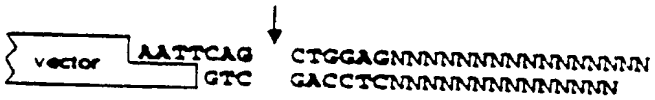


2. Transform library and isolate plasmid DNA

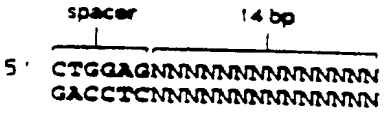
3. Digest DNA with BpmI



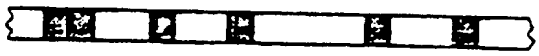
4. Digest DNA with PvuII



5. Isolate 20 bp **transcript markers** and treat with T4 DNA **polymerase**



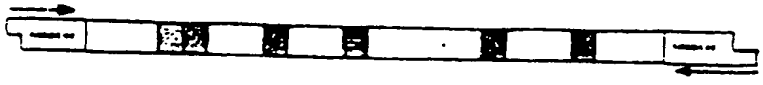
6. Concatenate blunt **markers** to form **arrays**



7. Add on adapters to ends of **marker arrays.**

FIGURE 5A

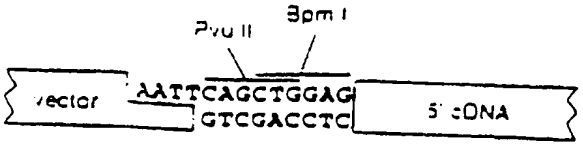
8. PCR amplify marker arrays using adapter sequences.



9. Clone arrays to create ATM library.

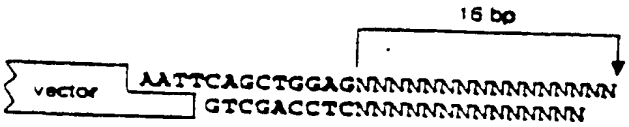
FIGURE 5B

1. Construct cDNA library using standard techniques with an EcoRI/BpmI adapter

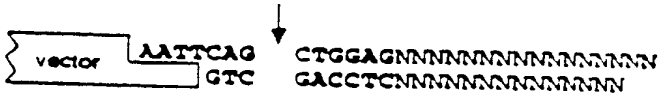


2. Transform library and isolate plasmid DNA

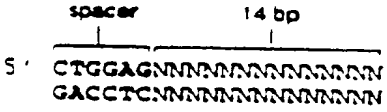
3. Digest DNA with BpmI



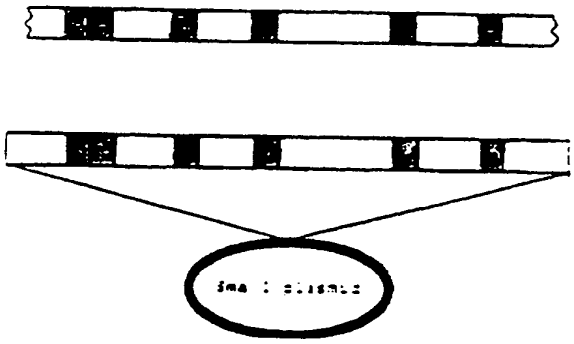
4. Digest DNA with PvuII



5. Isolate 20 bp **transcript markers** and treat with T4 DNA polymerase



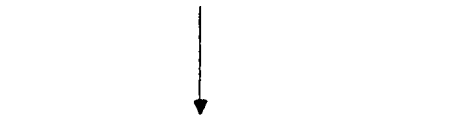
6. Concatenate blunt **markers** to form **arrays**



7. Ligate **arrays** into a plasmid vector.

FIGURE 6A

8. PCR amplify ligation with vector primers.



9. Clone arrays to create ATM library.

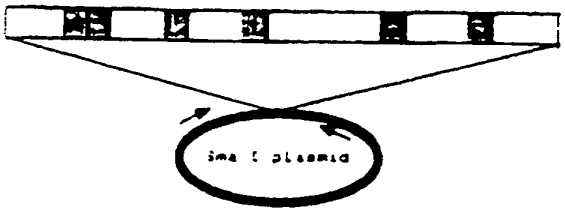
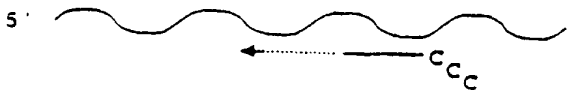
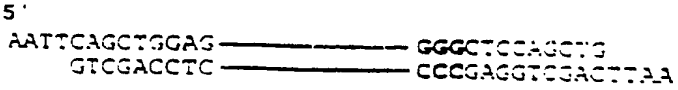


FIGURE 6B

1. Perform 1st strand cDNA synthesis using modified random hexamer

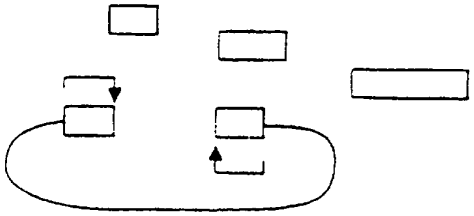


2. 2nd strand synthesis as usual; ligate Bpm I adapter



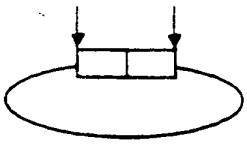
3. Clone into pSPORT-modified vector lacking Bpm I sites

4. Transform library and isolate plasmid DNA



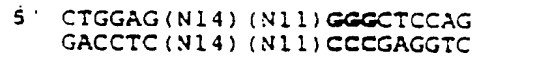
5. Digest DNA with Bpm I and T4 DNA polymerase treat

6. Dilute reaction, self-ligate, and re-transform library



7. Isolate plasmid DNA and digest with PvuII

8. Purify 40 bp transcript markers



9. Concatenate markers and clone arrays

FIGURE 7

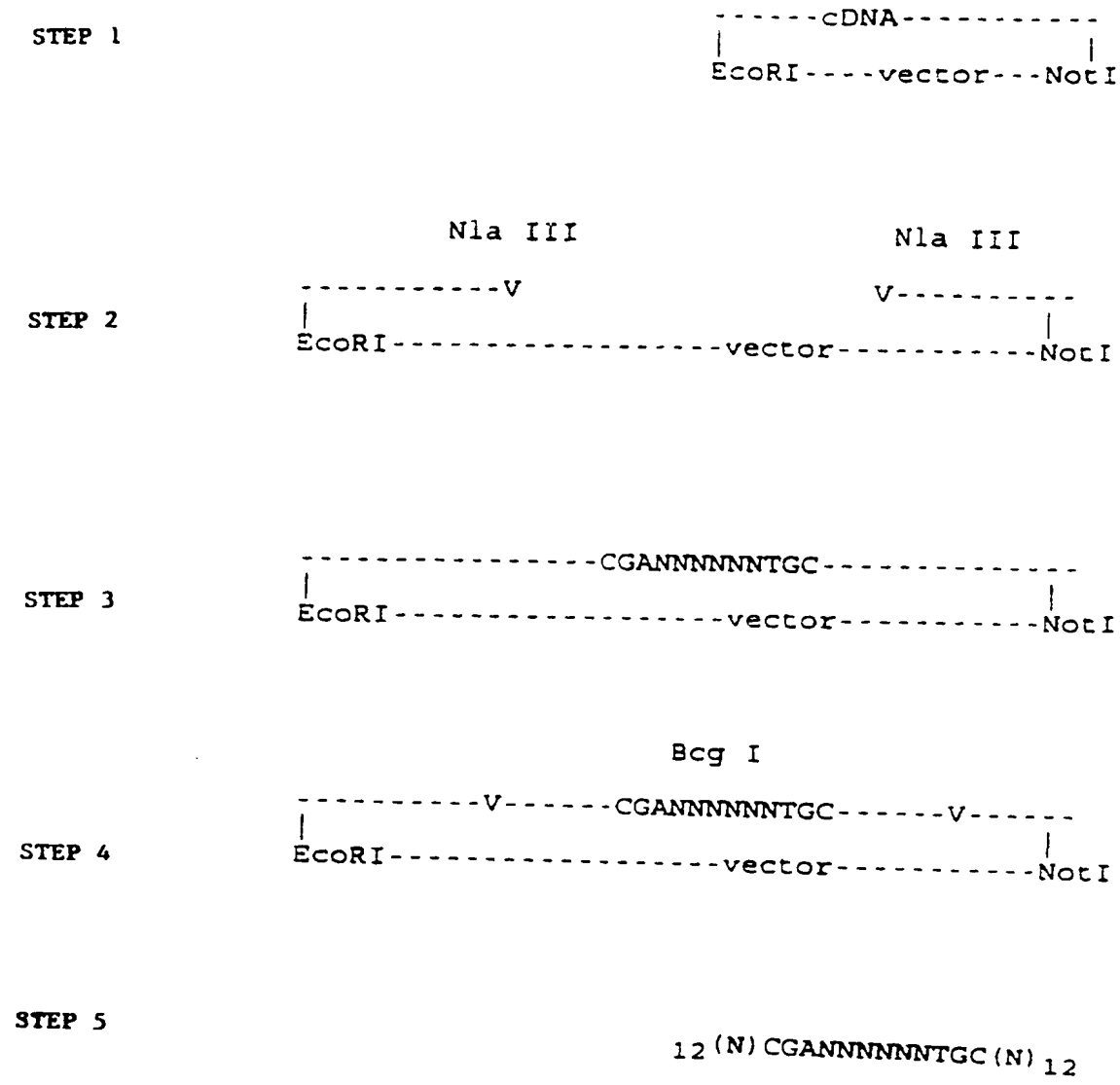


FIGURE 8

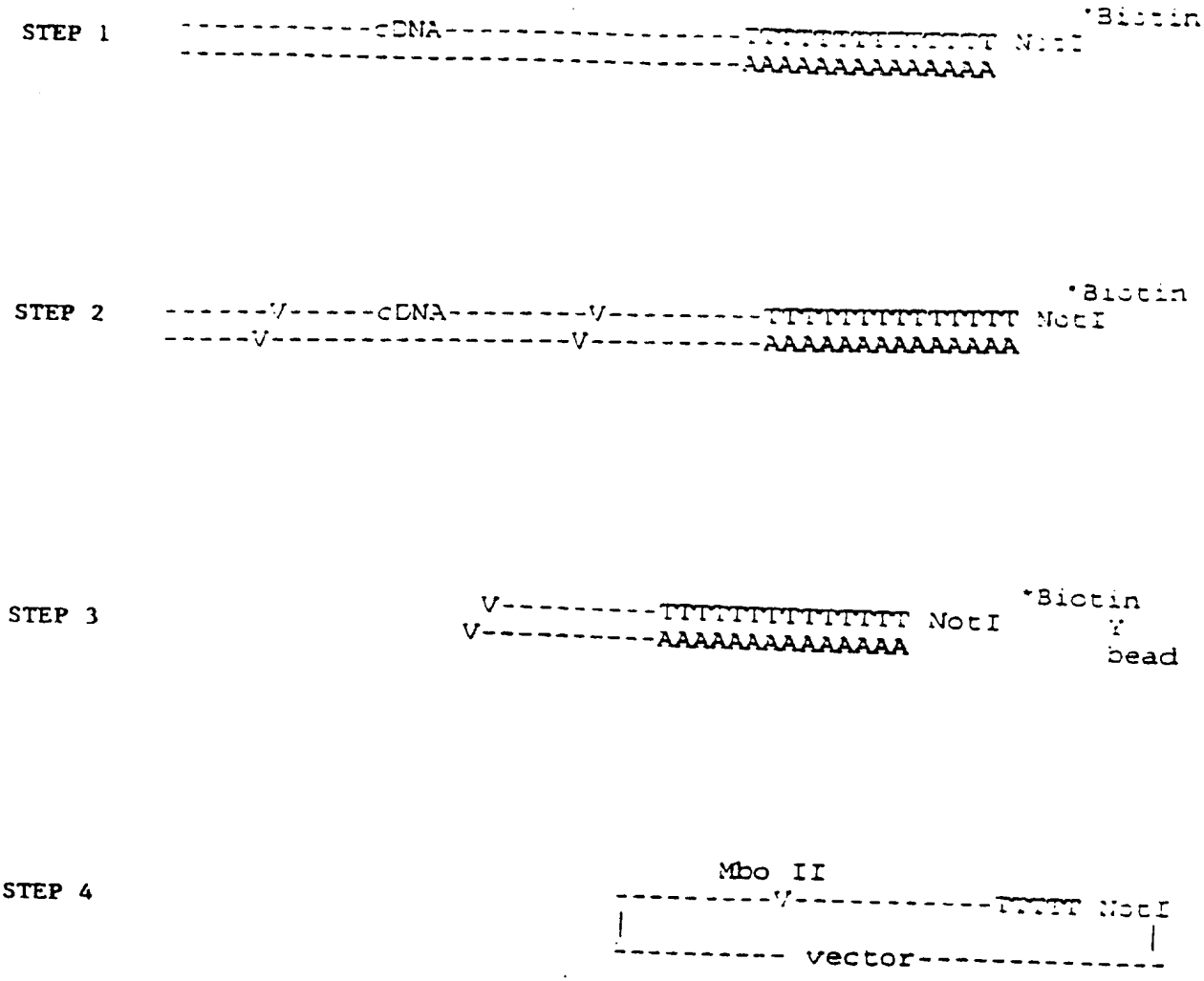


FIGURE 9

List of 12 known sequences

1. rat jagged protein/human Tislld gene/human mRNA (KIAA0068)
2. human lamin B receptor
3. human mRNA for calmodulin--PROSTUT04
4. human chromogranin A mRNA
5. rat mRNA for RNA binding protein/mouse mRNA for SIG-41-- PROSNOT06
6. human Ia-associated invariant gamma chain--PROSNOT02 and others
7. human chromosome X cosmid, clones 196B12--PROSTUT04
8. h. mac-2 binding protein mRNA- PROSNOT16
9. h. apoferritin H gene exon 1-PROSNOT11
10. h. HLA-B associated transcript 3-PROSNOT14
11. h. mitochondrial DNA/h. cortex mRNA containing Alu repetitive element
12. h. transcription factor ZFM1 mRNA-PROSNOT15

FIGURE 10

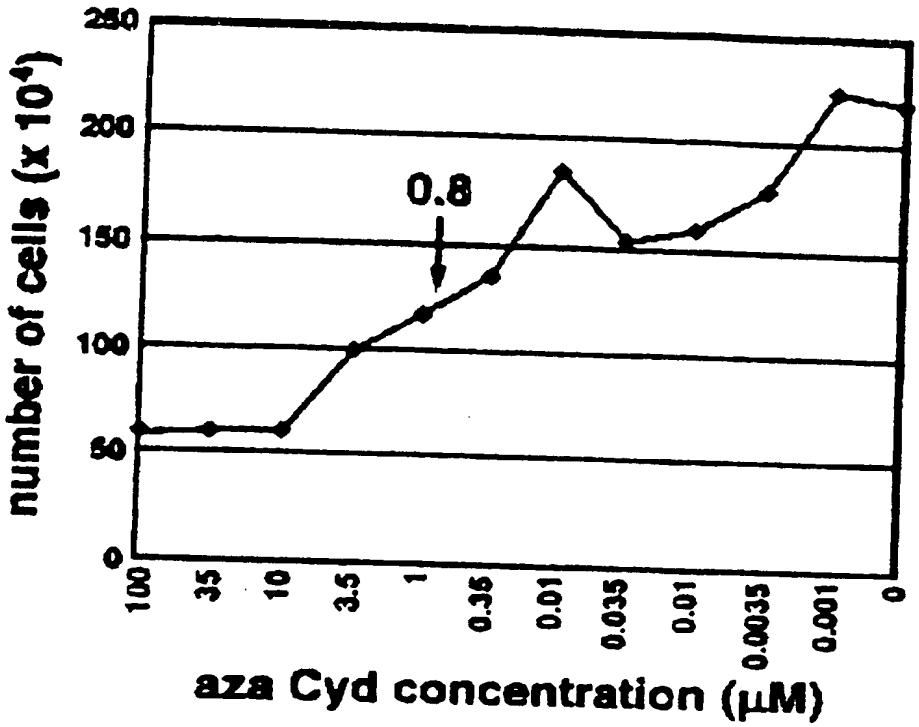


FIGURE 11

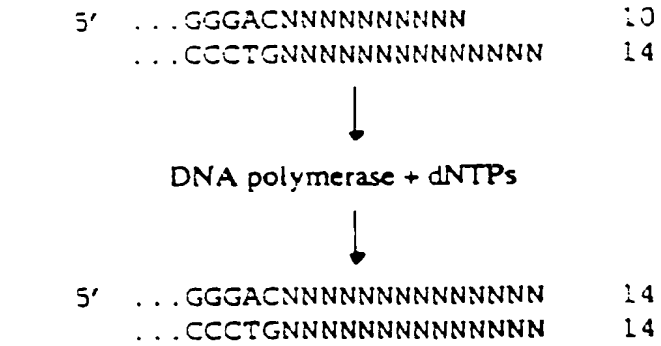


FIGURE 12A

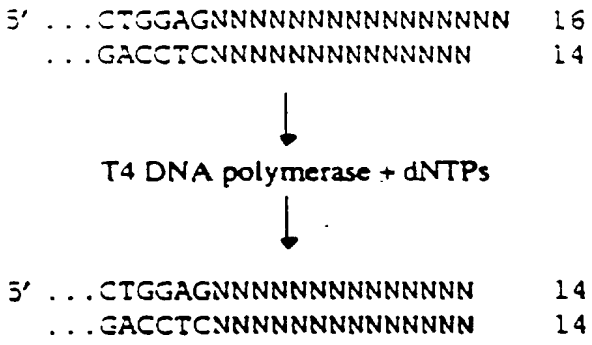
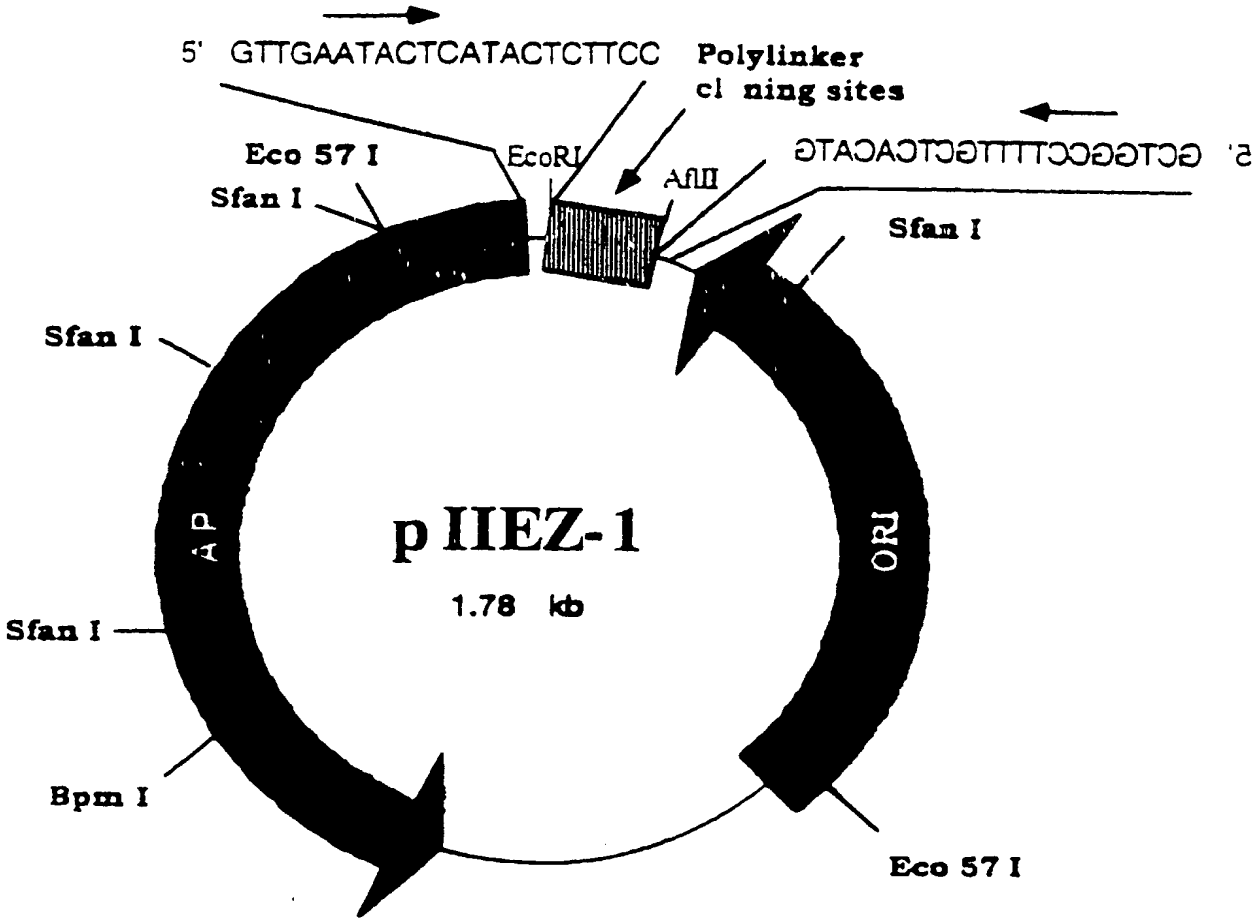


FIGURE 12B



10 20 30 40 50 60

GAATTCACCTGCAGGGACGTCTACTGAAGTACTTAAGGATGAATGCCTGGAGCTCGCATC

EcoRI BspMI BsmF I Aat II Rsa I FokI BpmI AluSfaN I
PstI Eco57 ISca I SacI
Sse8387I
Tai I

70 80

TGTGCAGCTGACGGCGCGCCGCGGCCGCA

Bsg I Alu AscI NotI
Bbv I
Pml I
PvuII

FIGURE 13

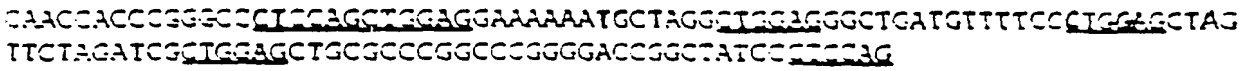


FIGURE 14A

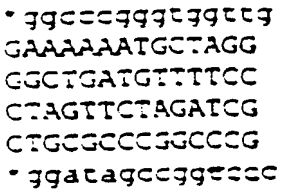


FIGURE 14B

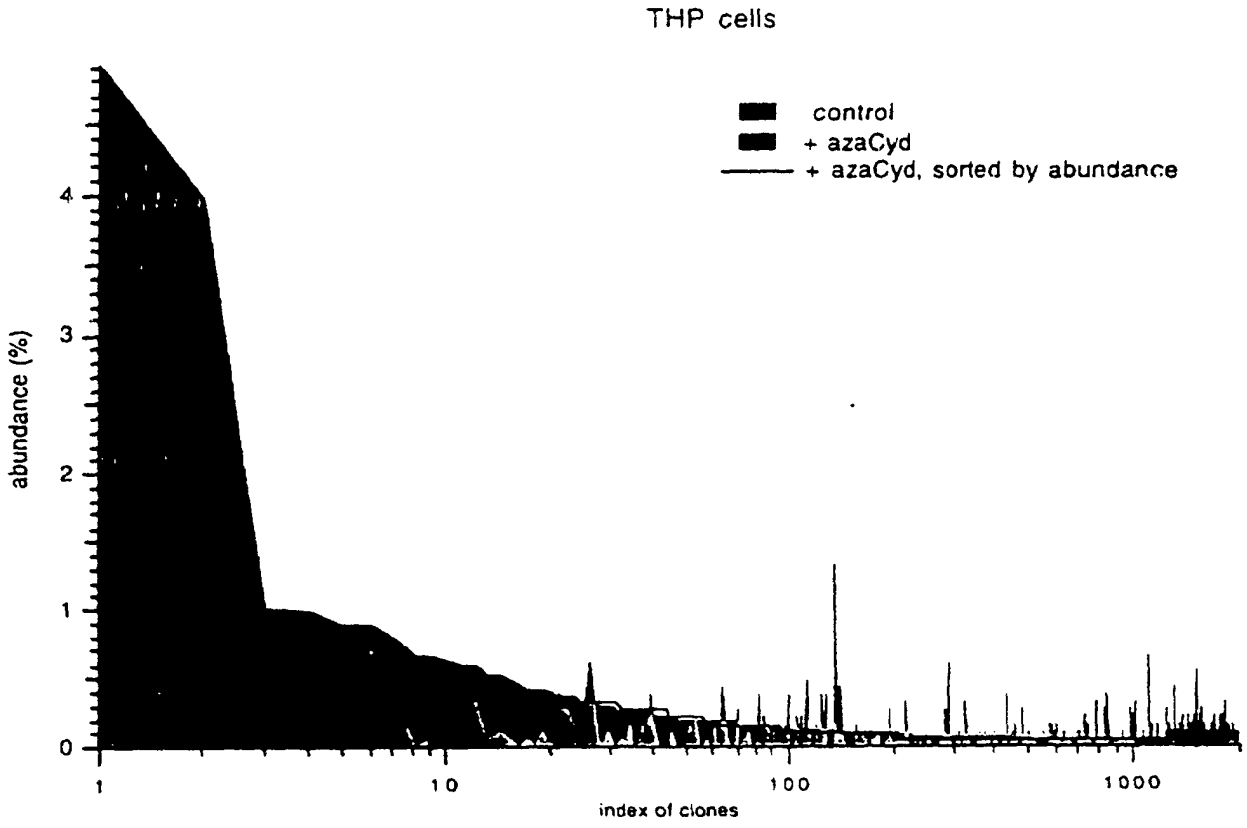


FIGURE 15

INTERNATIO SEARCH REPORT

[redacted]tional Application No
PCT/US 97/18344

A. CLASSIFICATION OF SUBJECT MATTER
IPC 6 C12Q1/68 C12N15/10

According to International Patent Classification(IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 6 C12Q

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document, with indication, where appropriate of the relevant passages	Relevant to claim No
X	KATO: "DESCRIPTION OF THE ENTIRE mRNA POPULATION BY A 3'END cDNA FRAGMENT GENERATED BY CLASS IIS RESTRICTION ENZYMES" NUCLEIC ACID RESEARCH, vol. 23, no. 18, 1995, pages 3685-3690. XP002053720 cited in the application see the whole document ---	1-48
X	US 5 508 169 A (DEUGAU KENNETH V ET AL) 16 April 1996 see the whole document ---	1-48

-/--

[X] Further documents are listed in the continuation of box C

[X] Patent family members are listed in annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

1

Date of the actual completion of the international search	Date of mailing of the international search report
28 January 1998	18/02/1998

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Hagenmaier, S

Form PCT/ISA/210 (second sheet) (July 1992)

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document with indication where appropriate of the relevant passages	Relevant to claim No.
X	VELCULESCU ET AL.: "SERIAL ANALYSIS OF GENE EXPRESSION" SCIENCE, vol. 270, 1995, pages 484-487, XP002053721 cited in the application see the whole document	1-48
A	WO 95 08647 A (UNIV COLUMBIA ;SOARES MARCELO B (US); EFSTRATIADIS ARGIRIS (US)) 30 March 1995 cited in the application see the whole document	5,23,33, 41
A	WO 95 20681 A (INCYTE PHARMA INC) 3 August 1995 see the whole document	6,7,15, 24,25, 42,43
A	JÜTTERMANN ET AL.: "TOXICITY OF 5-AZA-2'-DEOXYCYTIDINE TO MAMMALIAN CELLS IS MEDIATED PRIMARILY BY COVALENT TRAPPING OF DNA METHYLTRANSFERASE RATHER THAN DNA DEMETHYLATION" PNAS, vol. 91, 1994, pages 11797-11801, XP002053722 cited in the application see the whole document	8,9,26, 27,31, 32,44,45
A	US 4 994 370 A (SILVER JONATHAN ET AL) 19 February 1991 see the whole document	35

INTERNATIONAL SEARCH REPORT

Information on patent family members

International Application No
PCT/US 97/18344

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 5508169 A	16-04-96	CA 2036946 A	07-10-91
WO 9508647 A	30-03-95	US 5482845 A	09-01-96
		AU 7842594 A	10-04-95
		US 5637685 A	10-06-97
WO 9520681 A	03-08-95	AU 1694695 A	15-08-95
		BG 100751 A	31-07-97
		CA 2182217 A	03-08-95
		CN 1145098 A	12-03-97
		CZ 9602189 A	14-05-97
		EP 0748390 A	18-12-96
		FI 962987 A	26-09-96
		JP 9503921 T	22-04-97
		NO 963151 A	27-09-96
		PL 315687 A	25-11-96
		HU 75550 A	28-05-97
US 4994370 A	19-02-91	NONE	

CORRECTED VERSION*

PCT

WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 :
C12Q 1/68, C12N 15/10

A1

(11) International Publication Number: WO 98/14619

(43) International Publication Date: 9 April 1998 (09.04.98)

(21) International Application Number: PCT/US97/18344

(22) International Filing Date: 3 October 1997 (03.10.97)

(30) Priority Data:
08/723,646 3 October 1996 (03.10.96) US

(63) Related by Continuation (CON) or Continuation-in-Part (CIP) to Earlier Application
US 08/723,646 (CIP)
Filed on 3 October 1996 (03.10.96)

(71) Applicant *(for all designated States except US)*: INCYTE PHARMACEUTICALS, INC. [US/US]; 3174 Porter Drive, Palo Alto, CA 94304 (US).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: WANG, Bruce, B. [US/US]; 1123 Banyon Way, Pacifica, CA 94404 (US). CHUNG, Alicia [US/US]; 2939 20th Avenue, San Francisco, CA 94132 (US). GUEGLER, Karl, J. [CH/US]; 1048 Oakland Avenue, Menlo Park, CA 94025 (US). YANG, Zhi [CN/US]; 600 Coleman Avenue, Menlo Park, CA 94025 (US). COCKS, Benjamin, Graeme [AU/US]; 4292 D Wilke Way, Palo Alto, CA 94306 (US). STUART, Susan, G. [US/US]; 1256 Birch Street, Montara, CA 94037 (US).

(74) Agent: BILLINGS, Lucy, J.; Incyte Pharmaceuticals, Inc., 3174 Porter Drive, Palo Alto, CA 94304 (US).

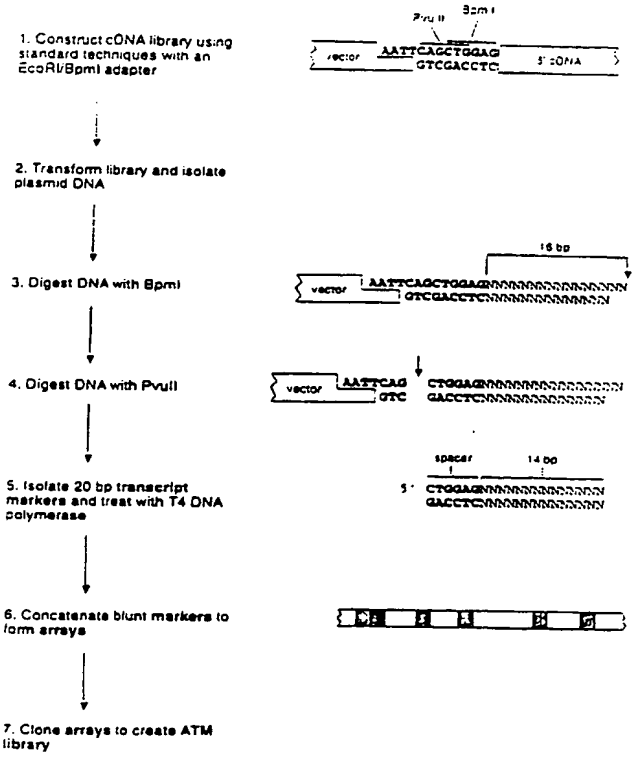
(81) Designated States: AT, AU, BR, CA, CH, CN, DE, DK, ES, FI, GB, IL, JP, KR, MX, NO, NZ, RU, SE, SG, US, ARIPO patent (GH, KE, LS, MW, SD, SZ, UG, ZW), European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, ML, MR, NE, SN, TD, TG).

Published
With international search report.
Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.

(54) Title: METHODS FOR GENERATING AND ANALYZING TRANSCRIPT MARKERS

(57) Abstract

The present invention relates generally to the field of molecular biology and specifically to rapid, high–throughput gene discovery methods that facilitate genome closure and to methods for analyzing gene expression patterns. The present invention provides methods and vectors useful for constructing libraries of transcript markers. The present invention also provides sequence specific methods for extending the nucleotide sequence of partial transcripts in a high–throughput manner using polymerase chain reaction.



*(Referred to in PCT Gazette No. 38/1998, Section II)

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania
AM	Armenia
AT	Austria
AU	Australia
AZ	Azerbaijan
BA	Bosnia and Herzegovina
BB	Barbados
BE	Belgium
BF	Burkina Faso
BG	Bulgaria
BJ	Benin
BR	Brazil
BY	Belarus
CA	Canada
CF	Central African Republic
CG	Congo
CH	Switzerland
CI	Côte d'Ivoire
CM	Cameroon
CN	China
CU	Cuba
CZ	Czech Republic
DE	Germany
DK	Denmark
EE	Estonia
ES	Spain
FI	Finland
FR	France
GA	Gabon
GB	United Kingdom
GE	Georgia
GH	Ghana
GN	Guinea
GR	Greece
HU	Hungary
IE	Ireland
IL	Israel
IS	Iceland
IT	Italy
JP	Japan
KE	Kenya
KG	Kyrgyzstan
KP	Democratic People's Republic of Korea
KR	Republic of Korea
KZ	Kazakstan
LC	Saint Lucia
LI	Liechtenstein
LK	Sri Lanka
LR	Liberia
LS	Lesotho
LT	Lithuania
LU	Luxembourg
LV	Latvia
MC	Monaco
MD	Republic of Moldova
MG	Madagascar
MK	The former Yugoslav Republic of Macedonia
ML	Mali
MN	Mongolia
MR	Mauritania
MW	Malawi
MX	Mexico
NE	Niger
NL	Netherlands
NO	Norway
NZ	New Zealand
PL	Poland
PT	Portugal
RO	Romania
RU	Russian Federation
SD	Sudan
SE	Sweden
SG	Singapore
SI	Slovenia
SK	Slovakia
SN	Senegal
SZ	Swaziland
TD	Chad
TG	Togo
TJ	Tajikistan
TM	Turkmenistan
TR	Turkey
TT	Trinidad and Tobago
UA	Ukraine
UG	Uganda
US	United States of America
UZ	Uzbekistan
VN	Viet Nam
YU	Yugoslavia
ZW	Zimbabwe